THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Comprehensive Analysis of Sustainable Flood Retention Basins

Qinli Yang

THE UNIVERSITY
of EDINBURGH

A thesis submitted for the degree of Doctor of Philosophy

School of Engineering

The University of Edinburgh

2011

Declaration

I declare that this thesis was completed by myself, that information derived from the work of others has been referenced, and that this work has not been submitted in any form for any other degree or professional qualification.

Qinli Yang

September 2011

Abstract

To adapt to climate change which results in increasing flood frequency and intensity, the European Community has proposed Flood Directive 2007/60/EC. It requires member states to conduct risk assessments of all river basins and coastal areas and to establish Flood Risk Management Plans focused on prevention, protection and preparedness by 2015. Sustainable Flood Retention Basins (SFRB) that impound water are a new concept that arose in 2006. They can have a pre-defined or potential role in flood defense and were supposed to facilitate the implementation of the Flood Directive. Early and preliminary studies of SFRB were derived from case studies in Southern Baden, Germany. In Scotland, there are a relatively high number of SFRB which could contribute to flood management control. This research aimed to produce a guidance manual for the rapid survey of SFRB and to propose a series of frameworks for comprehensive analysis and assessment of SFRB. Precisely 372 SFRB in central Scotland and 202 SFRB in Southern Baden were investigated and characterized by 43 holistic variables. Based on this practical experience, a detailed guidance manual was created, guiding users to conduct a SFRB survey in a standardized and straightforward way. To explore the hidden data structure of data arising from the SFRB survey, various widely used machine learning algorithms and geo-statistical techniques were applied. For instance, cluster analysis showed intrinsic groupings of SFRB data, assisting with SFRB categorization. Principal Component Analysis (PCA) was applied to reduce the dimensions of SFRB data from the original 43 to

23, simplifying the SFRB system. Self-organizing Maps (SOM) visualized the relationships among variables and predicted certain variables as well as the types of SFRB by using the highly related variables. Three feature-selection techniques (Information Gain, Mutual Information and Relief) and four benchmark classifiers (Support Vector Machine, K-Nearest Neighbours, C4.5 Decision Tree and Naïve Bayes) were used to select and verify the optimal subset of variables, respectively. Findings indicated that only nine important variables were required to accurately classify SFRB. Three popular multi-label classifiers (Multi-Label Support Vector Machine (MLSVM), Multi-Label K-Nearest Neighbour (MLKNN) and Back-Propagation for Multi-Label Learning (BP-MLL)) were applied to classify SFRB with multiple types. Experiments demonstrated that the classification framework achieved promising results and outperformed traditional single-label classifiers. Ordinary Kriging was used to estimate the spatial properties of the flood-related variables across the research area, while Disjunctive kriging was used to assess the probability of these individual variables exceeding specific management thresholds. The results provided decision makers with an effective tool for spatial planning of flood risk management. To assess dam failure hazards and risks of SFRB, a rapid screening tool was proposed based on expert judgment. It demonstrated that the levels of *Dam Failure Hazard* and *Dam Failure Risk* varied for different SFRB types and in different regions of central Scotland. In all, this thesis provided a guidance manual for rapid survey of SFRB and presented various effective, efficient and comprehensive frameworks for SFRB analysis and assessment, helping to promote the understanding and management of SFRB and thus to contribute to Flood Risk Management Plans in the context of the Flood Directive.

Publications

A: Water Resources Research

1. Journal papers

[1] **Yang, Q.**, Shao, J., Scholz, M., Böhm, C., Plant, C. and Tumulac, P.: Multi-label classification model for Sustainable Flood Retention Basins. (Submitted to Environmental Modelling & Software).

[2] **Yang, Q.**, Scholz, M. and Shao, J. (2011): Application of Spatial Statistics as a Screening Tool for Sustainable Flood Retention Basin Management. Water and Environment Journal, ISSN 1747-6585 (Accepted in May 2011).

[3] **Yang, Q.**, Shao, J., Scholz, M. and Plant, C. (2011): Feature selection methods for characterizing and classifying adaptive Sustainable Flood Retention Basins. Water Research 45(3), 993-1004.

[4] Scholz, M. and **Yang, Q.** (2010): Guidance on Variables Characterising Water Bodies including Sustainable Flood Retention Basins. Landscape and Urban Planning 98(3-4), 190-199.

[5] McMinn, W.R., **Yang, Q.** and Scholz, M. (2010): Classification and assessment of water bodies as adaptive structural for flood risk management planning. Journal of. Environmental Management 91(9), 1855-1863.

[6] Wang, W., Tang, X., Huang, S., Zhang, S., Lin, C., Liu, D.W., Che, H.J., **Yang, Q.** and Scholz, M. (2010): Ecological Restoration of Polluted Plain Rivers within the Haihe River Basin in China. Journal of Water Air Soil Pollution 211(1-4), 341-357.

2. Conference papers

[1] **Yang, Q.**, Shao, J. and Scholz, M. (2011): Classification of Water Bodies including Sustainable Flood Retention Basins (SFRB). International Conference on Integrated Water Resources Management (IWRM 2011). 12th-13th October, 2011, Dresden, Germany. (Accepted for oral presentation).

[2] **Yang, Q.**, Shao, J. and Scholz, M. (2010): Geostatistical Assessment of Wetlands that Can Be Used as Sustainable Adaptive Structural Measures for Diffuse Pollution and Flood Risk Management Planning. Proceedings of the 12th International Water Association International Conference on Wetland Systems for Water Pollution Control, Venice, 4-9 October 2010, Volume 2, 1581-1588 (in Italy).

[3] **Yang, Q.**, Shao, J. and Scholz, M. (2010): Assessment of the Classification of Sustainable Flood Retention Basins with a Self-organizing Map Model. Proceedings of the 12th International Water Association International Conference on Wetland Systems for Water Pollution Control, Venice, 4-9 October 2010, Volume 2, 1101-1108 (in Italy).

[4] **Yang, Q.**, McMinn, W.R. and Scholz, M (2010): Assessment and Classification of Scottish Water Bodies as Sustainable Adaptive Hydraulic Structures

for Flood Risk Management Planning. Proceeding of the 1st European IAHR Congress incorporating the 3rd International Junior Researcher and Engineer Workshop on Hydraulic Structures (IJREWHS '10). Edinburgh, 4-6 May 2010 (in United Kingdom).

[5] **Yang, Q.**, McMinn, W.R. and Scholz, M. (2009): Potential Use of Natural Flood Retention Wetlands to Control Diffuse Pollution. In J. M. Bayona and J. Garcia (Eds.), Proceedings (on CD) of 3rd Wetland Pollutant Dynamics and Control WETPOL 2009, Barcelona, 20-24 September 2009 (in Spain), 79-80.

B: Computer Science-Data Mining

1. Journal papers

[1] Shao, J., He, X., Böhm, C., **Yang, Q.** and Plant C.: Synchronization-inspired Partitioning and Hierarchical Clustering. (Submitted to IEEE Transactions on Knowledge and Data Engineering (TKDE)).

[2] Shao, J., Hahn, K., **Yang, Q.**, Wohlschläger, A., Böhm, C., Myers, N. and Plant, C (2010): Hierarchical Density-Based Clustering of White Matter Tracts in the Human Brain. International Journal of Knowledge Discovery in Bioinformatics 1(4), 1-26. **[Best Article for IGI Global's "Fourth Annual Excellence in Research Journal Awards"]**

[3] He, D., Shao, J., Geng, N. and **Yang, Q.** (2008): A Model for Image Categorization Based on Biological Visual Mechanism. New Zealand Journal of Agricultural Research 50(5), 781-787.

2. Conference Papers

[1] Shao, J., Plant, C., **Yang, Q.**, and Böhm, C. (2011): Detection of Arbitrarily Oriented Synchronized Clusters in High-dimensional Data (Accepted by International Conference on Data Mining (ICDM) 2011) (in Canada).

[2] Shao, J., Hahn, K., **Yang, Q.**, Böhm, C., Wohlschlaeger, A., Myers, N. and Plant, C. (2010): Combining Time Series Similarity with Density-based Clustering to Identify Fiber Bundles. Proceedings of International Conference on Data Mining (ICDM), Workshop on Biological Data Mining and its Applications in Healthcare, Sydney, 14-17 December 2010, 747-754 (in Australia). [**Best Paper Award**]

[3] Shao, J., Böhm, C., **Yang, Q.**, Plant, C. (2010): Synchronization Based Outlier Detection. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2010), Barcelona, 20-24 September 2010, 245-260 (in Spain).

[4] Böhm, C., Plant, C., Shao, J. and **Yang, Q.** (2010): Clustering by Synchronization. Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010), Washington DC, 25-28 July 2010, 583-592 (in USA).

[5] Shao, J., He, D. and **Yang, Q.** (2008): Multi-semantic Scene Classification Based on Region of Interest. Proceedings of International Conferences on Computational Intelligence for Modeling, Control and Automation (CIMACA), Vienna, 10-12 December 2008, 732-737 (in Austria).

The following above-mentioned papers have evolved from my doctoral thesis A1[1], A1[2], A1[3], A1[4], A1[5], A2[2], A2[3].

Acknowledgements

First and foremost, I would like to express my warmest thanks to my supervisor Prof. Dr. Miklas Scholz for his outstanding supervision, invaluable suggestions and endless kind support. Your initiative, enthusiasm and hardworking on research encouraged me to further develop as a scientific researcher. I am also grateful to Dr. Rory Harrington for assessing my first year report and gave me many useful suggestions.

I would also love to thank Prof. Dr. Chrisitian Böhm and Dr. Claudia Plant, for opportunities of collaboration, the expert guidance and the plenty of kind help. Learning from you, I gained much knowledge on data mining technology which is the essential part of my thesis.

I thankfully acknowledge the technical support given by Mr. William R. McMinn during the field work. I also thank the Scottish Water for data support. Gratitude also goes to the project students including Chris Sagar, Louise Blackhall, Monica Prosser, Laurie-Anne Maubayou, Mark Keane, Rodney Moody, Michelle Robinson, and Prince Osei Bonsu. I enjoyed working with them for field work and pleasant discussions.

Many thanks are due to Dr. Martin Edward Parkes, my foreign teacher and good friend. He has been encouraging and helping me with my study since 2007 when

I was a postgraduate student in China. I was inspired by many of his world-wide ideas, long-run perspectives and constructive comments. As time went on, I opened my mind. Also thank him for helping me with the grammar checking in the thesis.

Help in various forms received from administrators in the Graduate School, the Institute of Infrastructure and Environment and the IT service. Mrs Margaret Taylor, Mrs Liz. Paterson, Mrs Sue Simpson, Ms. Joan Birse, and Mr. David Stewart are gratefully acknowledged.

I would wish to acknowledge the China Scholarship Council and The University of Edinburgh (CSC-UoE joint Scholarship) for funding my PhD study. I also acknowledge the DAAD and The University of Edinburgh Development Trust for their financial support.

I am very thankful to many of my friends in Edinburgh for bringing me happiness and warmth. Particularly, much appreciation is given to my roommate Yuan Wang for her warm care, sincere accommodation and kind help. Deep thanks also go to Yu Dong who has the same supervisor with me, for his trust, good advice, generous help and interesting stories. I also would like to express my deep gratitude to Miss Alison Furber for her kind help with the final grammar checking.

Finally, I appreciate my family, especially my husband Junming Shao, from the bottom of my heart. They are always my powerful backing with their unselfish love, understanding and generous accommodation. Without their support, this thesis would not have seen its conclusion.

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

Flooding is the most common natural disaster in Europe [61]. It can endanger lives and livelihoods, and devastate public infrastructure and services, resulting in destructive and long lasting impacts on individuals, communities and businesses. Climate change predictions suggest that the magnitude and frequency of extreme precipitation events are likely to increase [60, 44], which may lead to more severe and frequent floods throughout Europe [100]. Therefore, this increase in amount of risk calls for a need to establish an effective, efficient and comprehensive framework for flood risk management.

In this chapter, the background information of the research, which includes the Flood Directive, flood defence measures and the novel concept of Sustainable Flood Retention Basins (SFRB) are first introduced in Section 1.1. Section 1.2 presents the aims and objectives of the thesis. Research contributions to knowledge are summarized in Section 1.3. Finally Section 1.4 outlines the structure of this thesis.

1.1 Background

1.1.1 Flood Directive

The PESETA (Projection of Economic impacts of climate change in Sectors of the European Union based on boTtom-up Analysis) research project [24] reported that under the 3.9°C and 2.5°C climate change scenarios, the 100-year return discharge levels were projected to increase in many parts of Europe. In accordance with Intergovernmental Panel on Climate Change (IPCC) predictions for other parts of the world, the number and size of floods are likely to increase across Scotland. The Hadley Centre regional climate model experiments for Scotland indicate an annual precipitation increase of around 20% by 2080 - 2100 [71]. While, the rainfall depth in many other areas of the UK will increase 20-40% by 2080s [189]. The increase of the intensity and frequency of floods would threaten people's lives and result in devastating damage. A significant increase in people exposed to floods would occur mainly in the central European regions and the British Isles, where people would suffer increases in expected damage. The PESETA project report [24] showed that river flooding would affect 250,000 to 400,000 additional people per year in Europe by the 2080s, more than doubling the number in the period of 1961-1990. Between 7.7 billion and 15 billion € would be the total additional cost of damage from river floods in the 2080s, which is more than twice as much as the annual average cost over the 1961-1990 period [24].

The European Union (EU) has thus responded to an increase in the perceived severity of flooding by introducing the Flood Directive 2007/60/EC for the assessment and management of flood risks [39]. The EU member states are required to first identify the river basins and coast lines at risk of flooding by 2011. Then they would need to map the flood risks and hazards by 2013 for the

risk areas. Finally, they are expected to establish Flood Risk Management Plans by 2015 to take adequate and coordinated measures to reduce flood risk.

The Flood Directive shall be coordinated with the Water Framework Directive (WFD) [38]. WFD aims to establish a framework for water protection and improvement across Europe, requiring to manage Europe's water environment in an integrated and sustainable way. WFD is implemented based on the idea of river basin management, requiring to produce comprehensive river basin management plans that can help to deliver good water quality for each river basin [127]. One of the WFD's objectives is to mitigate the flooding effects, however, it is not explicitly addressed. In 2003, the role of the WFD in flood protection was widely recognized at European level on the Precautionary Flood Protection in Europe International Workshop [127]. It suggested that the implementation of the WFD could be regarded as a 'window of opportunity' to strength sustainable flood protection. Therefore, the WFD provides a framework to integrate flood risk management into river basin management, which indicates that Flood Directive should be carried out in coordination with the WFD. Furthermore, Member States shall take into consideration long term developments, e.g. climate change and sustainable land use practices in the flood risk management cycle addressed in the Flood Directive.

1.1.2 Flood Defence Options

Although it will never be possible to eradicate flooding, a wide range of traditional flood defence approaches were used to reduce the devastating impacts of flooding in urban and rural areas. State-of-the-art measures and technologies for coping with floods can be briefly classified into four groups: indigenous, non-structural, structural and holistic approaches [56]. In general, structural measures are

designed to reduce flood probability while non-structural measures are intended to reduce potential flood damage.

The indigenous method coping with floods advocates that humans live with floods wisely. Since flooding is a natural hazard and unavoidable, the attitude of respecting floods and doing nothing or very little to mitigate flooding is reasonable. This adaption approach can be found throughout South-East Asia and is still practiced in some areas [56]. According to this maxim, people do not settle in the flood-prone lands but use those lands for agriculture or grazing. Plants in the flood plains usually play roles in reducing soil erosion and improving deposition of fine sediments, which bring nutrients to the soil. In Netherland, since 2006, the project of **Room for the River** has been implemented involving 17 partners. It is a government design plan aims to increase safety by protecting the rivers region from future floods and improve the environmental quality in the areas surrounding Holland's rivers [46]. This strategic policy has been made by training and regulating river channels and thus minimizing the damage of flooding. In the future it could be selected when an overwhelming environmental interest arises.

Non-structural measures typically refer to measures designed for reducing the impacts and consequences of floods without changing flood properties [43]. They consist of three categories: regulation, defence, and flood insurance [192]. These measures vary from emergency response planning and training, raising public awareness, flood forecasting and warning, land use regulation to flood proofing and flood insurance [5, 183]. Humans have a passive attitude towards plans to floods, namely focusing on the adaptation of the socio-physical environment to the residual flood damage [190]. Nowadays, this kind of measure is receiving increasing attention.

As a basis of flood protection, structural measures will remain essential options

for flood control, especially in extreme flooding situations. Currently, five classic structural methods are: storage in reservoirs; storage in parts of the floodplain or other flat lands in the lower river reaches; improvement of river channel(s); creation of additional flood ways (so-called bypasses) and flood embankments [183]. One issue of these traditional engineered structures is that they are always expensive to construct and maintain. Another issue is that during heavy rainfall, this kind of artificial system can reach breaking point. Consequently, the surface water can overwhelm the existing sewer system and exceed the capacity at sewage treatment plants. This results in Combined Sewer Overflows (CSOs) where untreated sewage is discharged into rivers or the sea directly, which might seriously pollute rivers and/or sea. For example, in June 2009, heavy rains washed storm sewage and high levels of pollutants from the streets of Birmingham into tributaries of the River Trent, which is one of the largest rivers in England, causing the death of thousands of fish [177]. Moreover, enlarging or increasing these 'hard engineering' structures to adapt to climate change would require high cost.

Holistic approaches are the combinations of diverse alternative structural and non-structural measures. They may be the most attractive way of coping with floods due to the specificity and complexity of any specific catchment. However, complicated multi-criteria methods need to be used to select the appropriate measures by considering economics, environmental conversation, safety of people and property.

More recently, a more sustainable approach, Sustainable Urban Drainage Systems (SUDS) was proposed and used in Europe [191]. Examples of SUDS include constructed wetlands, ponds, grass filter strips and other porous surfaces designed to take overflowing water from impervious roads and drains. Operating in a natural manner, SUDS stored or absorbed water and slowed the runoff from urban areas, attempting to reduce flow peak and associated flooding damage. However

these approaches are only applicable at an urban/local scale and are not suitable at a catchment/large scale.

1.1.3 Sustainable Flood Retention Basins (SFRB)

To mitigate flood risk on a large scale, more innovative adaptive approaches are needed. In this context, the concept of Sustainable Flood Retention Basins (SFRB) is proposed, supplementing SUDS to ensure effective sustainable flood risk management planning. A SFRB is defined as an impoundment or integrated wetland, which has a pre-defined or potential role in flood defence and diffuse pollution control that could be accomplished cost-effectively through best management practice, supporting sustainable flood risk management and enhancing sustainable drainage, pollution reduction, biodiversity, green space and recreational opportunities for society. Examples of SFRB are basins used to supply electricity generating stations, lochs, reservoirs, integrated wetlands, even the water bodies in public parks, etc. Utilizing the available storage capacity of natural rivers or the existing reservoirs/basins to retain flood water and then release water or infiltrate water to achieve flood reduction, less new concrete defences and sustainable drainage systems would need to be built, and the combined sewage system would receive less runoff requiring storage and subsequent treatment. Generally, SFRB can be regarded as an adaptive, sustainable and cost-effective approach to mitigate flooding risks.

Facing significant challenges in complying with the EU Flood Directive, EU member states have financed Interreg IVB projects under the North Sea Region Programme (2007-2013). One of the projects is the Strategic Alliance for integrated Water Management Actions [155], which aims to adapt existing water management systems to the effects of extreme flood events, focusing on sustainable development of society and regional economies. The project comprises

22 partner institutions from Germany, Netherlands, Sweden, United Kingdom and Norway and integrates local, regional and national stakeholders, universities and vocational training students. As one part of the SAWA project, the work in this thesis focuses on Sustainable Flood Retention Basins, aiming to develop guidance and tools for adaptive measure SFRB to assist the member states in developing flood risk management plans.

1.2 Aims and Objectives

As a whole, this inter-disciplinary research aims to produce a detailed guidance manual for rapid survey of Sustainable Flood Retention Basins (SFRB) and to propose various effective frameworks to comprehensively analyze SFRB by applying machine learning algorithms and geo-statistics, which provide effective, reliable and efficient tools for engineers, scientists, authorities, decision-makers to better understand, assess and mange SFRB. The specific objectives are further listed as follows:

- To determine and characterize relevant variables of SFRB;
- To develop a guidance manual for a rapid and comprehensive method for surveying water bodies including SFRB;
- To find the intrinsic groups/classes of SFRB based on Cluster analysis;
- To reduce the dimensionality of the SFRB system descriptors by using Principal Component Analysis (PCA);
- To visualize the correlations among variables of SFRB, to predict the difficult-to-determine variables, and to predict the types of SFRB by using the Self-organizing Map (SOM) technique;

- To explore robust feature selection methods to achieve variable selection and further to verify the efficiency of selected variables for optimal classification;
- To classify SFRB using multi-label classification learning algorithms, allowing one SFRB belongs to more than one type simultaneously;
- To do spatial analysis of flood related variables of SFRB aiming to provide adaptive spatial solutions to flood risk mitigation;
- To assess the dam failure of SFRB across the study area;
- To use extensive case studies in Scotland and Baden to verify the effectiveness of the proposed methods;

1.3 Contributions

In this thesis, I comprehensively analyze SFRB by using with different techniques, e.g. data mining, machine learning and geostatistics. This research has two main components: (a) SFRB investigation and (b) SFRB analysis. The first component highlights how to characterize SFRB and how to survey SFRB with a rapid method, aiming to establish SFRB data sets. The second component deals with the idea of exploring the hidden data structure underlying the data sets by using diverse machine learning algorithms and Geo-statistics. The major contributions of this thesis can be summarized as follows:

1. It develops an integrative framework for SFRB study, consisting of conception, function, characterization, investigation, and categorization of SFRB. The basic elements of the framework are a set of 43 characteristic variables, 7 primary purposes of SFRB and 6 target types of SFRB. Field survey can be carried out by using the 43 variables to characterize each individual SFRB and referring its purposes to judge SFRB functions and types.

2. It offers a guidance manual for the rapid survey of SFRB. The guidance manual describes each characteristic variable in detail and tells the reader how to use the manual to assess SFRB. As a reference and benchmark, the guidance manual helps to train new users to implement SFRB survey work in a user-friendly way. Moreover, it keeps the SFRB assessment consistent.

3. It formulates an effective and meaningful clustering framework of SFRB. Based on the whole set of 43 variables, the SFRB data is automatically clustered into different groups, which correspond to different SFRB types. It helps to identify the intrinsic groups of SFRB and thus to match them with the six types of SFRB.

4. Dimensionality reduction of SFRB data is achieved by using Principal Component Analysis (PCA). The multivariate SFRB data is projected into low dimensions in a new space, simplifying the SFRB system. The importance of the variables is ordered according to their contributions to the principal components.

5. An effective and intuitive way to visualize the correlations among 43 variables is proposed by applying the Self-Organizing-Map technique. Based on the correlations, the SOM model can predict the difficult-to-determine variables by using their highly related variables. Furthermore, the types of SFRB can also be predicted by the SOM model.

6. It provides a simple, rapid and effective framework for variable selection and SFRB classification. Three feature selection techniques (Information Gain, Mutual Information and Relief) are applied on the SFRB dataset to identify the importance of the variables in terms of classification accuracy. Four traditional classifiers (Support Vector Machine, K-Nearest Neighbours, C4.5 Decision Tree and Naïve Bayes) are subsequently used to verify the effectiveness of the classification with the selected variables and automatically identify the optimal number of variables. Only nine important variables

are sufficient to accurately classify SFRB. The findings help to reduce the redundancy and complexity of the SFRB system and improve the further classification scheme.

7. It provides an effective, efficient and comprehensive framework for SFRB multi-label classification, allowing one SFRB to belong to more than one class, which will be helpful to better assess the real functions and status of SFRB. Three popular multi-label classifiers: Multi-Label Support Vector Machine (MLSVM), Multi-Label K-Nearest Neighbour (MLKNN) and Back-Propagation for Multi-Label Learning (BP-MLL) are applied to predict the types of SFRB based on two data sets (one from Scotland and another from Baden). Experiments demonstrate that the proposed classification framework achieved promising results and outperformed traditional classifiers.

8. An effective screening tool for water engineers for flood control using SFRB has been provided. Ordinary kriging is applied to estimate numerical values for all key flood control variables everywhere in the central Scotland. Moreover, the probability that certain threshold values relevant for flood control managers are exceeded can also be calculated by using Disjunctive kriging. Finally, kriging maps are produced for better management and spatial planning of SFRB for flood risk reduction.

9. A new flexible, rapid and affordable procedure for dams failure assessment of SFRB is proposed. *Dam Condition* along with the corresponding *Dam Failure Hazard* and *Dam Failure Risk* are estimated, analyzed and graphed. This preliminary assessment is the elementary step for the implementation of Flood Directive, which helps to identify the risk zones.

10. All the achievements stated above about SFRB analysis and assessment give guidance for Europe Union member states to comply with the Flood Directive. The deep insight into SFRB and better understanding of the real and

complicated functions of SFRB help to manage and develop SFRB properly for flood risk reduction and adaptation to climate change. Therefore, the SFRB research contributes to developing Flood Risk Management Plans.

1.4 Outline of the Thesis

The content of this thesis is organized as follows.

Chapter 1 provides the background of flood management in a very general manner to present the reader with the broader context of this thesis.

Chapter 2 is dedicated to an overview of Sustainable Flood Retention Basins (SFRB). It starts by introducing various aspects surrounding SFRB, which include the SFRB's concept, SFRB typology, locations of research area, the characterizing variables of SFRB and the primary purposes of SFRB. Afterwards, the procedure of data acquisition and the description of two available SFRB data sets are given.

Chapter 3 focuses on the guidance manual for the rapid survey of water bodies including sustainable flood retention basins. The guidance manual explains the characteristic variables of SFRB and how to evaluate these variables.

Chapter 4 presents the main techniques for SFRB analysis, which include clustering, classification, principal component analysis (PCA), self-organizing map (SOM), feature selection and geo-statistics. For each approach, the rationale, related work, theoretical explanations and application on real data are provided respectively.

Chapter 5 presents the research results and discussion. In addition, the applications of some methods on the case studies are also justified.

Chapter 6 gives the conclusions of the thesis. It outlines the achievements of this research as well as the limitations. Finally, the future research directions are guided.

Chapter 2

SUSTAINABLE FLOOD RETENTION BASINS

In this chapter, various basic components of SFRB are described. First, the origin of the SFRB concept is traced in Section 2.1. The suggested six types of SFRB are then described in Section 2.2. The identification of SFRB sites in research areas are illustrated in Section 2.3. The variables used to capture the characteristics of SFRB are discussed in Section 2.4. Section 2.5 briefly gives the primary purposes of SFRB. In Section 2.6, the procedure of data collection and the SFRB data sets that have been built up are described.

2.1 Concept

The concept of Sustainable Flood Retention Basins (SFRB), which was first proposed by Scholz [158], was derived from research on retention basins in the River Rhine Valley, Southern Baden, Germany. A SFRB was originally defined as an aesthetically pleasing retention basin, predominantly used for

flood protection while adhering to sustainable drainage and best management practices [159]. With the increased experience of SFRB in central Scotland, the previous definition was no longer sufficient. So it was revised and upgraded. A SFRB is defined as an impoundment or integrated wetland, which has a pre-defined or potential role in flood defence and diffuse pollution control that could be accomplished cost-effectively through best management practice, supporting sustainable flood risk management and enhancing sustainable drainage, pollution reduction, biodiversity, green space and recreational opportunities for society. The word 'sustainable' in SFRB means capable of being maintained at a steady level without exhausting natural resources, harming the environment or causing severe ecological damage.

2.2 SFRB Typology

The SFRB have diverse functions. For example, most SFRB are used for the collection of river flow and runoff and are emptied slowly resulting in downstream discharge waves being flattened and discharge periods extended, mitigating potential flooding [159, 160]. Some retention basins perform additional tasks such as infiltration for ground water recharge, drinking water supply, diffuse pollution mitigation, enhancement of recreational benefits such as water skiing, bird watching and fishing, and green space provision. In fact, some SFRB have even become Sites of Special Scientific Interest (SSSI) after years of neglect, which has resulted in a high biodiversity. Moreover, one basin often has multiple and mixed functions. Therefore, it is essential to assess the functions of SFRB properly by recognizing both its design purpose and current uses. Otherwise, it might lead to conflicts between stakeholders over the status and function of an SFRB and hinder successful Flood Risk Management Plans implementation. Therefore, it is highly necessary to determine the exact type of each SFRB, which will largely help

planners, engineers, local authorities and community groups better understand and communicate the status and functions of SFRB.

The SFRB typology supplements other wetland classification systems. Hydrological, geo-morphological, chemical and biological factors were used in a hierarchical classification systems of wetland and deepwater habitats located in the United States of America. Five major wetland types (Lacustrine, Riverine, Palustrine, Marine and Estuarine) were generated. These were further subdivided into more specific categories [101]. Furthermore, the Ramsar Classification of Wetland Types, which is based on physical and limnological characteristics, divides wetlands into three main categories (marine and coastal wetlands, inland wetlands and man-made wetlands) and 43 associated wetland types [139]. However, these classification systems are complicated, and detailed data sets, which are frequently unavailable, are required. Moreover, these methods do not accommodate the diverse roles of SFRB. Therefore, to appropriately meet the practical need to identify the main functions of the basins, in this study, six general SFRB types were suggested by an international group of civil engineers, landscape planners and environmental scientists [159]. Specifically, SFRB are classified as Hydraulic Flood Retention Basin (Type 1), Traditional Flood Retention Basin (Type 2), Sustainable Flood Retention Wetland (Type 3), Aesthetic Flood Treatment Wetland (Type 4), Integrated Flood Retention Wetland (Type 5) and Natural Flood Retention Wetland (Type 6). For illustration, Table 2.1 gives the detail description of each type in the context of SFRB research in Scotland.

Table 2.1: Definitions of Sustainable Flood Retention Basin (SFRB) types.

Type	Name	Definition of SFRB type	Typical examples	Characteristics
1	Hydraulic Flood Retention Basin (HFRB)	Managed, current SFRB that is hydraulically optimized and captures sediment	Basins used to feed electric stations; highly engineered and large flood retention basin	Very high rainfall and seasonal impact; high site elevation; normal floodplain elevation; very highly engineered or even automated with high outlet flexibility; fully managed and tidy in appearance; very high flood water volume; very deep flooding depth; potentially high basin gradient; very large flood surface area; very long wetted perimeter and dam; very problematic terrestrial and aquatic animal passage; often high levels of algae in spring and summer; can be permanently wet or virtually permanently dry; inorganic sediment; if purpose built SFRB often very high pollution
2	Traditional Flood Retention Basin (TFRB)	Traditional retention basin used for flood protection adhering to sustainable drainage and best management practice	Former drinking water reservoir; traditional flood retention basin	Very high rainfall and high seasonal impact; high site elevation; managed, highly engineered or even automated with high outlet flexibility: quite high flood water volume and deep flooding depth; potentially high basin gradient; large flood surface area and long wetted perimeter; high and long dam: problematic aquatic and terrestrial animal passage: algal bloom in summer; mostly inorganic sediment; low vegetation cover; not excessively polluted; little groundwater infiltration; mixed catchment with forestry and farming

Continued on next page

Table 2.1: (continued)

Type	Name	Definition of SFRB type	Typical examples	Characteristics
3	Sustainable Flood Retention Wetland (SFRW)	Retention and treatment wetland used for passive flood protection adhering to sustainable drainage and best management practice	Sustainable drainage systems or best management practices such as some retention basins, detention basins, large ponds or wetlands	High rainfall and clearly recognizable seasonal impact; relatively low score for engineered appearance; some outlet flexibility: acceptable aquatic and terrestrial animal passage; small to medium flood water volume and typically shallow flooding depth; normal for dam height and length: average wetted perimeter and flood surface area; usually highly polluted if wet; partly wet; mainly inorganic substrate (construction); if matured and unmanaged, the sediment becomes increasingly organic; substantial vegetation cover; average (highly urbanised) catchment size
4	Aesthetic Flood Treatment Wetland (AFTW)	Aesthetically pleasing treatment wetland for the retention and treatment of contaminated runoff, which is well integrated into the landscape and has some social and recreational benefits	Some modern constructed treatment wetlands; integrated constructed wetland	Fairly low rainfall; highly engineered with emphasize on water treatment; high flood water volume and shallow flooding depth: acceptable aquatic and terrestrial animal passage; flat and short dam; short wetted perimeter; large flood surface area; usually highly polluted and often wet; substantial and often lush vegetation; average catchment size with mixed uses
5	Integrated Flood Retention Wetland (IFRW)	Integrated flood retention wetland for passive treatment of runoff; flood retention and enhancement of recreational benefits	Some artificial water bodies within parks or near motorways that have a clear multi-purpose function such as water sport and fishing	Semi-natural; flat and short dam; low flood water volume and very shallow flooding depth; small flood water surface area and short wetted perimeter; easy animal passage if not in urban areas; usually highly polluted with high organic proportion in sediments; usually substantially wet; very high, dense and lush vegetation cover; small catchment (often like a park)

Continued on next page

Table 2.1: (continued)

Type	Name	Definition of SFRB type	Typical examples	Characteristics
6	Natural Flood Retention Wetland (NFRW)	Passive, natural flood retention wetland that may have become a site of Special Scientific Interest (SSSI) requiring protection from adverse human impacts	Natural or semi-natural lakes and large ponds, potentially with restricted access	Very natural and most likely a site of specific scientific interest (SSSI) or at least with a high potential for a SSSI; dam typically absent and no outlet flexibility; very low flood water volume and very shallow flooding depth; often very small flood surface area (unless a large managed lake) and short wetted perimeter; easy aquatic and terrestrial animal passage; usually very wet or permanently wet; usually deep; natural organic sediment (originating predominately from basin vegetation for mature SFRB); little pollution; very high proportion of vegetation cover; very small catchment with dominant pasture cover; high groundwater infiltration; possibly neglected for decades; high rather natural catchment proportions

To intuitively illustrate the different types of SFRB, representative pictures for typical examples of SFRB types are further exhibited in Figure 2.1 - Figure 2.6 as follows.

Figure 2.1: Lednock Reservoir ($56.43°N$, $4.08°W$) is a typical example of a Hydraulic Flood Retention Basin (SFRB Type 1).

Figure 2.2: Glensherup Reservoir ($56.22°N$, $3.67°W$) is a typical example of a Traditional Flood Retention Basin (SFRB Type 2).

Figure 2.3: Garnqueen Loch ($55.89°N$, $4.05°W$) is a typical example of a Sustainable Flood Retention Wetland (SFRB Type 3).

Figure 2.4: Dundas Loch ($55.97°N$, $3.41°W$) is a typical example of an Aesthetic Flood Treatment Wetland (SFRB Type 4).

Figure 2.5: Beveridge Park ($56.10°N$, $3.17°W$) is a typical example of an Integrated Flood Retention Wetland (SFRB Type 5).

Figure 2.6: Lindores Loch ($56.33°N$, $3.19°W$) is a typical example of a Natural Flood Retention Wetland (SFRB Type 6).

2.3 Identification of SFRB Locations

In this study, 372 and 202 sites were selected and located for SFRB study using the 1:50 000 scale Ordnance Survey Maps of central Scotland (Figure 2.7) and Southern Baden, Germany (Figure 2.8), respectively. Alternatively, all of these SFRB sites could be located on Google Earth or a digital map based on their available geographical coordinates. In some cases, the sites, which were newly built up and with features not shown on the maps, were discovered during the field work. In the context of this investigation, the sites of interest were those which might be able to play a role in either flood management or diffuse pollution control. Structures that might be able to play a role in flood control were considered to be those where the water level could be controlled either manually or automatically, and are typically former or current engineered water supply reservoirs. Sites with the potential to contribute to diffuse pollution control were typically more natural and relatively small water bodies. Specifically, in relation to dam failure risk assessment of SFRB, the interesting sites were those which had a dam that could play a role in flood control.

It is noticeable that the distribution of SFRB differs in the two survey areas. Specifically, SFRB are sparsely and evenly located in central Scotland while SFRB in Souther Baden are more relatively concentrated near city areas. The reason behind this phenomenon might be: many SFRB in central Scotland are either private (used for finishing, water sports, etc.) or natural water bodies, most of which are situated in upland or remote areas; In contrast, a large number of SFRB in Southern Baden are purpose built basins used for flood protection and thus they are centralized near cities. Briefly, these two figures reflect the difference of design approaches between central Scotland and Southern Baden.

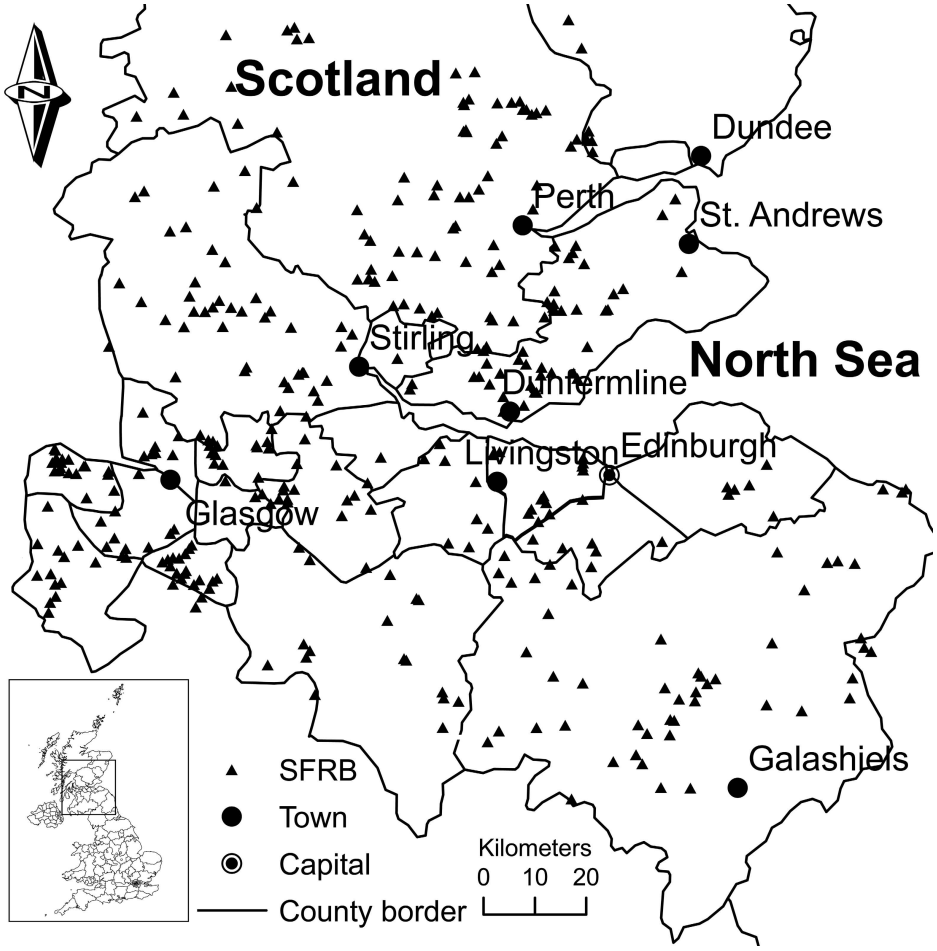


Figure 2.7: Study area, administrative boundaries and the 372 identified sustainable flood retention basins (SFRB) in central Scotland (United Kingdom).

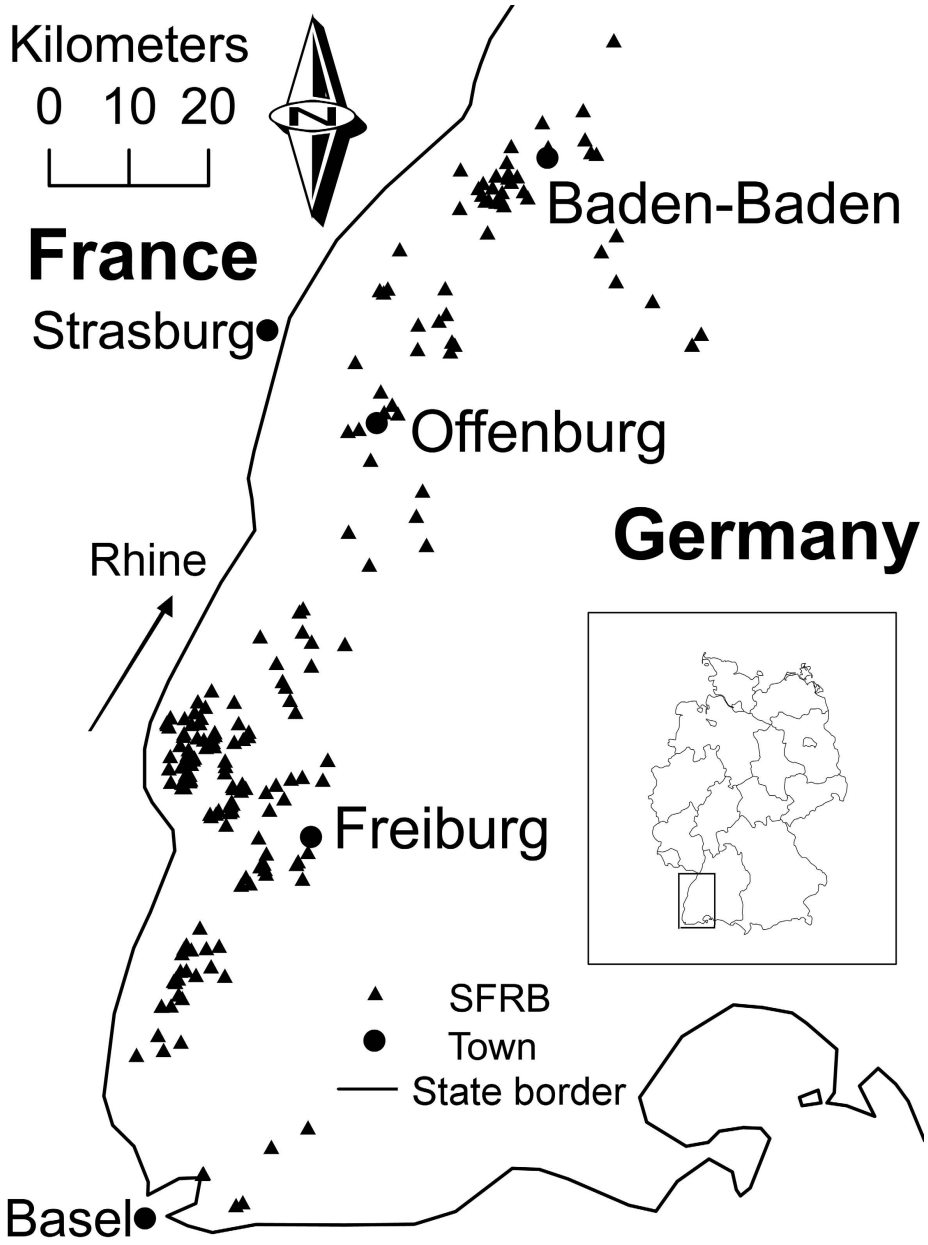


Figure 2.8: Study area, administrative boundaries and the 202 identified sustainable flood retention basins (SFRB) in Southern Baden (Germany).

2.4 Characterizing Variables of SFRB

Based on the literature review, various site visits and group discussions among European engineers, scientists, environmentalists and landscape and urban planners, Scholz [159] initially identified 34 variables characterizing retention basins in Southern Baden. To fit the SFRB framework in Scotland, all the original 34 variables were adjusted and extended. Currently, 43 variables have been developed to capture the characteristics of SFRB. Each specific variables as well as its ID are listed in Table 2.2. The newly proposed variables are marked in bold.

The current 43 variables were developed based on the following rules.

Supplement: The new variables *Estimated Flood Duration*, *Mean Depth of Basin*, *Length of Basin* and *Width of Basin* were introduced in the context of Scotland water bodies. In addition, to assess the risk at the SFRB, three new variables (including *Dam Condition*, *Dam Failure Hazard*, and *Dam Failure Risk*) and their components were proposed to supplement the existing 40 variables.

Split: The previous variable *Aquatic and Land Animal Passage* was divided into the following separate variables: *Aquatic Animal Passage* and *Land Animal Passage*. This accounts for fundamentally different obstacles concerning the freedom of unrestricted movement for animals. Similarly, the previous variable *Forest and Natural Catchment Proportion* was split into *Forest Catchment Proportion* and *Natural Catchment Proportion*.

Refine: The variables such as *Engineered*, *Floodplain Elevation*, *Basin and Channel Connectivity*, *Mean Flooding Depth*, *Estimated Flood Duration*, *Wetness*, *Typical Wetness Duration*, *Wetted Perimeter* *Vegetation Cover* and *Relative Total Pollution* were refined to fit within the Scottish context. It has been noticed that there are differences in the built environment and landscape. For example,

Table 2.2: Characteristic variables used for the assessment of sustainable flood retention basins.

ID	Variable and unit	ID	Variable and unit
1	Engineered (%)	23	Site Elevation (m)
2	Dam Height (m)	24	Vegetation Cover (%)
3	Dam Length (m)	25	Algal Cover in Summer (%)
4	Outlet Arrangement and Operation (%)	26	Relative Total Pollution (%)
5	**Aquatic Animal Passage (%)**	27	Mean Sediment Depth (cm)
6	**Land Animal Passage (%)**	28	Organic Sediment Proportion (%)
7	Floodplain Elevation (m)	29	Flotsam Cover (%)
8	Basin and Channel Connectivity (m)	30	Catchment Size (km^2)
9	Wetness (%)	31	Urban Catchment Proportion (%)
10	Proportion of Flow within Channel (%)	32	Arable Catchment Proportion (%)
11	Mean Flooding Depth (m)	33	Pasture Catchment Proportion (%)
12	Typical Wetness Duration (d/yr)	34	Viniculture Catchment Proportion (%)
13	**Estimated Flood Duration (d/yr)**	**35**	**Forest Catchment Proportion (%)**
14	Basin Bed Gradient (%)	**36**	**Natural Catchment Proportion (%)**
15	Mean Basin Flood Velocity (cm/s)	37	Groundwater Infiltration (%)
16	Wetted Perimeter (m)	**38**	**Mean Depth of Basin (m)**
17	Maximum Flood Water Volume (m^3)	**39**	**Length of Basin (m)**
18	Flood Water Surface Area (m^2)	**40**	**Width of Basin (m)**
19	Mean Annual Rainfall (mm)	**41**	**Dam Condition (%)**
20	Drainage (cm/d)	**42**	**Dam Failure Hazard (%)**
21	Impermeable Soil Proportion (%)	**43**	**Dam Failure Risk (%)**
22	Seasonal Influence (%)		

the variable *Mean Flooding Depth* recognizes high slope values for the Scottish landscape and deep flooding depths of some natural lakes. The variable *Wetness* was further refined to make a strong distinction between permanently wet systems

such as reservoirs and lakes (Scottish data set) and SFRB, which might be dry and became wet only occasionally (German data set).

Remove: The original variable *Flood Frequency*[90] was deleted since its estimation required some information which was not always available or inapplicable, such as peak discharge. The former variable *Viniculture Catchment Proportion* was not suitable for Scotland, therefore, it was also removed.

As these variables combined hard scientific and engineering data e.g. the dimensions of dams and spillway, with softer more holistic landscape and environmental variables, they provide a comprehensive assessment of SFRB characteristics.

2.5 Primary Purposes of SFRB

As stated above, each SFRB often performs multiple functions. During the site visit in Southern Baden [159], five purposes of SFRB were noticed: hydraulics, sustainable drainage, environmental protection, recreational activities and landscape aesthetics. Relying on field work study in central Scotland, they were updated and enlarged, consisting of flood protection, water storage, sustainable drainage, environmental protection, recreation and landscape enhancement, industrial production (mainly of historic importance), drinking water supply and economic use (such as farming).

2.6 Data Acquisition

In this study, the data collection of each SFRB consists of a desk study and a field survey. During this process, 43 variables (see Table 2.2) and the main purposes of

the SFRB were assessed. Specially, the two stages of data acquisition are further described as follows.

2.6.1 Desk Study

The desk study provided an estimate of most variables by searching for relevant information from institutions, publications and digital databases etc. in usually 30 minutes to 1 hour. The desk study was supported by information obtained from the Scottish Flood Defence Asset Database [10] and other relevant regional references such as Foster et al.[49], Meteorological Office [79], Scottish Water, and Scottish Environmental Protection Agency. During the desk study, the catchment boundaries for the water body, land use of the catchment (urban, arable, forestry, and natural grassland proportion), wetted perimeter, area of the water body, length of the dam, elevation and basin gradient were measured, using 1:50,000 or 1:25,000 digital maps.

2.6.2 Field Work

The site visit, typically requiring between 30 and 120 minutes per site (depending on the experience of the team and the complexity of the SFRB), was required to verify the attributes determined during the desk study and the water body inflows and outflows were documented. Details regarding variables concerning the presence of a potential dam, its outlet control operation, basin catchment proportions, vegetation cover and drainage were documented during a site visit and by taking photographic evidence. The 43 attributes included conventional hard engineering variables such as *Dam Length* and *Dam Height*, along with more holistic variables such as how *Engineered* the structure appeared and variables such as *Aquatic Animal Passage* and *Land Animal Passage*. This combination of

hard and soft variables readily lent itself to solving multi-disciplinary problems such as sustainable water management. A guidance manual on how to determine the 43 variables characterizing water bodies including SFRB is provided in Chapter 3. A quality index (i.e. low = 1 to 40%; medium = 40 to 60%; high = 60 to 100%) was attributed to each variable during the desk and field studies to reflect the likelihood of selecting a correct value. Data with low quality index were subsequently improved by a further literature review or field revisits.

2.6.3 Data Sets

In central Scotland, a total of 372 SFRB were visited during October 2008 and March 2010 by various team members who were engaged in environment science and engineering, civil engineering, biology, chemistry and water resources management. During July to September 2010, 180 SFRB in Southern Baden were revisited and another 42 new sites were identified and surveyed during the field work. In total, 372 Scottish SFRB and 202 Baden SFRB were investigated, each of which held data for 43 variables and 7 primary purposes. Associated with each value of the variables, a quality index was assigned. In the following chapters, all experiments are based on the two data sets.

2.7 Summary

In this chapter, an overview of Sustainable Flood Retention Basins (SFRB) was presented. The basic concept, types and 43 characteristic variables of SFRB were introduced. These definitions provided a basic knowledge of SFRB, which is crutial for further SFRB analysis in Chapters 4 and 5. At the end of this chapter, data acquisition was further described.

Chapter 3

GUIDANCE MANUAL

The assessment and implementation of the concept of sustainable Flood Risk Management Plan is an emerging challenge in environmental and water management. The concept is being further advanced by research at The University of Edinburgh (cf. Chapter 4 and Chapter 5). It is recommended that these findings should be consulted prior to implementing Flood Risk Management Plan methodologies.

This chapter is dedicated to the guidance manual for the rapid assessment of SFRB. The guidance manual explains the underlying philosophy behind the rapid survey system (Section 3.1), introduces how to use the manual (Section 3.2) and provides advice on determining the variables for water bodies including SFRB in the field (Section 3.3). Section 3.4 discusses how to assess the primary purposes of SFRB. This part has been published on the journal of Landscape and Urban Planning.

3.1 Background

3.1.1 Rational for Rapid Survey Method

Existing survey methods for assessing the full status of water bodies are mainly based on the ecological, chemical and quantitative criteria. For example, under the Water Framework Directive [38], all surface and ground waters must aim to achieve "good ecological status" by 2015. Different criteria can be used for assessing the status of different waters (e.g. rivers, lakes, transitional waters and coastal waters), but they all need to consider the biological, physical and chemical quality, environmental quality (standards for levels of specific pollutants) and hydromorphological quality of the water bodies [132]. The methods to determine these characteristics are time consuming and expensive [187]. As these methodologies are predominantly ecological, their outputs tend to over-emphasize the ecological status of a water body, and this can give rise to conflict in the case of flood defence impoundments with high flood return periods, because these basins become overgrown and often achieve a high biodiversity (see Figure 3.1).

In some cases in Europe, this has resulted in expensive flood defence structures, which cannot be used because flooding would damage the ecology [74]. Many SFRB including reservoirs in Scotland and elsewhere lie immediately upstream of or adjacent to heavily populated areas. Flooding from reservoirs can result from an uncontrolled breach of the dam or overtopping during severe rainfall. This may result in catastrophic consequences for life, property, critical infrastructure and economy [42]. Such conflicts need to be resolved through impartial debate and discussion with an objective assessment of the structure, its design purpose and current status. Many existing hydrological models do not consider the flood control potential of existing dams and impoundments to contribute to hydraulic management, though reservoir release from drinking water reservoirs to maintain

river ecology is an established management practice [121]. Pitt [131] recommended the creation of inundation maps particularly of those areas near reservoirs. This recommendation has been implemented for most areas of the country [42, 166]. Some of the more critical information of interest for emergency services includes flood extent, water depth, flood water velocity, hazard level, time of initial inundation and time of peak arrival. The shortcomings of this assessment are that the determination of most of these variables is very costly and not very accurate. Moreover, the total complexity and process dynamic can never be fully captured, and changes over time. It follows that there is a need for a rapid and cost-effective assessment of risk-related variables.

Figure 3.1: Long return period sustainable flood retention basin becoming overgrown and developing tree cover and marsh ecosystem in Southern Baden [158] .

A further aspect of sustainability in flood risk management should be to consider the existing flood defence infrastructure and impoundments that already exist within a catchment. Considering that many agencies will have to undertake

assessments of their areas and objectively classify the flood defence potential of the existing infrastructure [167], a detailed expensive investigation is not always going to be practical. The system outlined in this guidance manual has proven to be inexpensive, rapid and reliable as an assessment tool for existing flood retention basin infrastructure such as most SFRB. The SFRB concept has evolved since 2006, and is based on the views of diverse international groups of engineers, landscape planners and environmental scientists, and has withstood detailed scientific scrutiny.

3.1.2 Manpower and Equipment Requirements

The philosophy behind this methodology is that it is rapid and inexpensive to apply within a catchment, and therefore should not require expensive equipment or detailed measurements. The solution is a two stage process of combining a desk study and a field visit. The desk study can provide an estimate of most variables using a standard personal computer with an internet connection in about 1 hour. The site visit involves locating the water body, recording SFRB and catchment details using a digital camera, and assessing the SFRB variables visually. This typically requires 30 to 120 minutes per site. Five or more sites can be assessed within a day of fieldwork, and the data gathered can be fed into the SFRB assessment tool (e.g. [163]) to objectively categorize the surveyed structures.

A crucial feature of the proposed approach is that it should preferably be used by a multi-disciplinary group of assessors. Ideally, the group should have different areas of expertise such as engineering, environmental science, hydrology, landscape planning and/or flood control planning. A team of two to four is ideal, as it promotes discussion and debate during the surveys. It is possible to apply the method with a single assessor; however, there is often a risk that the outcome of the assessment is biased towards their particular discipline. The methodology has

been demonstrated to be most effective when different disciplines are combined within the assessment team. The team can be separated into two small groups, each of which fills in a survey template (see 3.1.3) independently and then discuss together to get an agreed assessment.

The basic equipment therefore required for the entire process is a personal computer with an internet connection to carry out research, a digital camera to record details of the catchment and the SFRB, and a 1:25000 to 1:50000 scale map of the survey area. A Global Positioning System (GPS) unit with $< 5m$ error is a useful additional tool to allow geographical locations to be recorded, which can subsequently be used in digital modeling. A range of receivers are also available for the European Magellan GPS system, and typically these can achieve accuracies of between 1 and 2 m with post processing modules [54]. The higher the quality of information used in the process, the more reliable the outcomes will be.

3.1.3 Survey Template

The SFRB survey method is based on completing the site survey template (Appendix A), which contains a total of 43 variables. Details of these are provided in the subsequent section with practical guidance on how to determine each of them. A vital aspect of the classification system is to assign an estimated confidence level to each of the variables as they are determined. The quality index (%) is an estimate by the assessors of how accurately each variable has been determined and the confidence that they have in the determination.

The confidence value has been banded into high, medium and low confidence levels. A high quality index is typically one that has been measured or can be estimated with a very high degree of confidence based on knowledge and

experience. The quality index then assigned is between 61% and 100% (Appendix A). In cases where the quality index is between 41% and 60%, additional investigations should be conducted to improve the quality index. In cases where the quality index assigned to a variable is $\leq 40\%$ the variable should be treated as missing. It has been found that assessors who use the system tend to assign quality index in 5% increments. In theory, 1-9 scale is suggested by AHP [150].

In addition to the section on site characterization variables, comprising details of the land types within a catchment and details of the SFRB and its hydraulics, there is a further section on "Assessment of the Primary Purposes of SFRB" (Section 3.4), which should be completed after the site visit. This section is potentially more subjective than the 43 basin variables, and considers what the structure has evolved into and its current range of uses, and therefore taking into consideration the sustainability of the structure while recognizing the SFRB design purpose.

The survey template is only a guide for case studies in temperate and oceanic climates. It requires modification to be applied effectively in other climatic zones and to accommodate various scales of infrastructure. In particular, descriptions for variables such as annual rainfall and seasonal impact should be adjusted to the application area. Moreover, it is recommended that national weather and mapping data should be used where available to decide on landscape and climatic variables, and the appropriate ranges for these parameters. As the assessment tool depends on the selection of the relative positions rather than on the numerical boundaries of the basins (Appendix A), it provides a relative objective and consistent output, which can be used to facilitate stakeholder discussions and identify infrastructure, which has the potential to be used in Flood Risk Management Plan.

3.2 How to Use the Manual

This manual contains all the required information to allow for a comprehensive water body (particularly potential SFRB) assessment in a survey case study area. It is intended that the method should be applied at a catchment scale where the catchment boundaries are defined and the SFRB characteristics within the catchment are then identified.

Once the locations of potential SFRB in a study area have been identified, a desk study for each basin is undertaken, using all available sources of information. Variables are then estimated or measured and recorded on the survey form (Appendices A and B). The next step is to perform a site visit to confirm the measured and estimated variables from the desk study and the survey form is completed and finalized. A short description of each of the variables is supplied in Section 3.3 of this guidance manual with information on how to determine each of the parameters. The entire process can be completed within approximately one hour for most sites.

Appendix A provides a description of the different variables and their boundary conditions within the remit of the survey. The classification boundaries have been made as widely applicable as possible. However, it is probable that some variable descriptions in the template will need to be updated to address fundamental differences in, for example, rainfall and terrain, when the methodology is applied for areas very different to Scotland and Baden, Germany. Appendix A can be printed out and laminated and used as a key for the variables in the field.

Appendix B provides a simplified survey form that should be printed out and used to gather the desk study information, and can then be filled in after the site visit. It is intended to be printed as a double sided A4 page and provides ample space to record additional information.

3.3 Assessment of the Characteristic Variables

3.3.1 Engineered (%)

A body of water can either be formed naturally or it can be created by man. Man-made structures can be highly diverse and range from very large water supply and hydropower dams of earth or concrete through to small scale structures often built to supply water to industrial processes (e.g. Figure 3.2 and Figure 3.3). The Engineered variable determines whether the SFRB is natural or man-made and how pronounced these tendencies are.

Figure 3.2: Highly engineered Sustainable Flood Retention Basin (Type 1).

Dam structures with full engineering control, such as a drinking water supply reservoir (Figure 3.2) or a purpose built HFRB are examples of potentially highly engineered structures. These types of basin would typically receive values of between 70% and 95%.

Figure 3.3: Natural Flood Retention Wetland (Type 6).

A structure that is natural is one that has not been extensively modified by man. The basin is typically a natural landscape feature and there is little (e.g. formal outlet or protected embankment) or no evidence of human interference (Figure 3.3). This type of basin will normally receive a value of <20%.

Section 3.4 on primary purposes has a strong influence on this variable, particularly in cases where the original design purpose and current use are different. Numerous industrial impoundments still exist in many areas, long after the industry they supported closed. These basins have become overgrown and neglected, often resulting in high biodiversity and in other cases, these basins are now designated nature reserves [82]. These industrial relics originally built for flood control now fulfil an entirely different role, and this is recognized in the division of scores between the different purpose categories.

3.3.2 Dam Height (m)

Dam height is defined as the height of the man-made structure, which creates the impoundment. The height of the structure is taken from the highest to the lowest point, usually just below the bottom outlet.

In some countries, databases of dams are available, which contain details of dam structures and reservoir capacities. Where this type of resource is available, it is recommended that it is utilized to gather accurate information for the survey sites.

For a rapid assessment, the dam height can be visually estimated, and a valuable tool is a photograph with the face of the dam and a team member on top. The team member can then be used to scale the structure. Accurate GPS can also be used to record spatial co-ordinates at the top and base of the SFRB, though these can have high levels of error associated with them as well. Typically, the height has to be estimated from the front face of the dam as this is clearly visible.

3.3.3 Dam Length (m)

The dam length is defined as the span of the structure creating the impoundment. The structure may span a valley or, in rare cases a dam may surround virtually the entire water body. Note that a natural lake restricted only by the topography of a valley attracts a dam length value of 0 m.

If written documentation of dam details is available, then this variable can be determined from this information. Dam length can be paced or measured during the site visit, or for structures, which are $> 500m$ long, they can be measured from maps with a scale of $\leq$ 1:50000.

3.3.4 Outlet Arrangement (%)

The outlet variable describes how and where water leaves the SFRB. In the case of small natural water bodies, there is usually a single river leaving the water body. This arrangement can be considered as a single, independent, simple and uncontrolled outlet. The river outlet will not have any form of dam, weir or sluice to control water levels and would generally receive a value of between 0 and 8% (with zero applied where there are no control structures). An entry of 8% might be given if there was a simple control or measurement structure present, but which does not significantly impact on water levels (Figure 3.4).

Figure 3.4: Simple gate at the outlet of a sustainable flood retention basin that does not restrict water flow.

In the case of Natural Flood Retention Wetlands (NFRW), the river outlet is typically natural and would receive a value of zero or close to it. The outlets in many shallow wetland systems can be choked by reeds and other vegetation, and

Figure 3.5: Heavily vegetated outlet of a Natural Flood Retention Wetland (Type 6).

these may slow down water release and treat it at the same time. This is one of the reasons why these natural systems are considered SFRB (Figure 3.5).

Many relatively small dams and impoundments have a minimum of two types of outlets. Typically, a dam will have an overflow (spillway) and pipes within the dam structure for various purposes; e.g. bottom outlet to drain the reservoir, outlet to control the river base flow and outlet to convey the water for subsequent treatment. A complex outlet structure is usually considered to be a combined system (typically with manual operation and a single fixed spillway) and attracts values between 15% and 75% (Figure 3.6). Higher values being awarded to highly engineered systems with potentially fully automated and remotely controlled structures.

Large and modern drinking water reservoirs or hydropower dam constructions

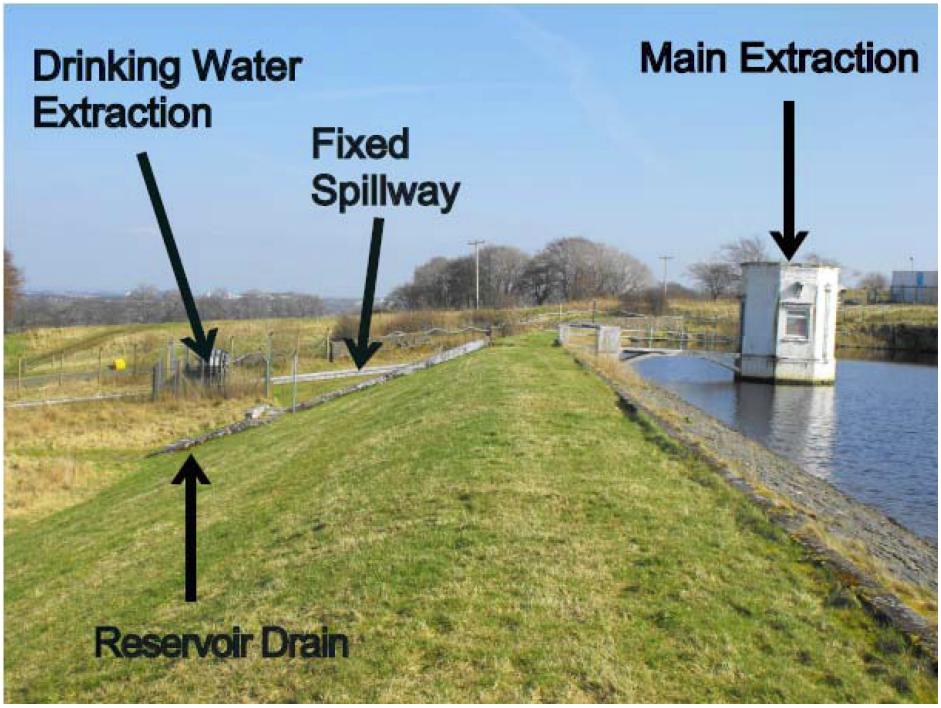


Figure 3.6: An example of a combined outlet with a fixed spillway and water abstraction system; 75% was assigned to the variable outlet arrangement.

typically have combined outlets with fixed spillways and one or more water extraction points. These larger dams are often fully automated and typically receive a value higher than 75%.

3.3.5 Aquatic Animal Passage (%)

Key to the movement of aquatic animals is the flow of water from the basin and the geomorphology of the water course or overflow. In the case of many dams, the only possible route of movement to upstream areas for fish and other aquatic organisms would be up the spillway and outlet pipes. Typically, these are long concrete or masonry channels with significant drops in height and very thin layers of water making aquatic animal passage almost impossible [98]. In the absence of

Figure 3.7: Dam spillway which would be a barrier to aquatic animal movement.

a fish pass or fish ladder, these should be considered significant barriers to aquatic animal passage, typically receiving values between 0 and 10% (Figure 3.7).

In the case of some smaller SFRB, steps may or may not have been taken to facilitate aquatic animal passage. Generally, those SFRB with an adequate flow of water, no significant drops in height and an insignificant barrier such as a small dam would score between 10% and 40%. Most SFRB that have been designed with a small fish pass or bypass stream would score between 41% and 69%.

Semi-natural water bodies and dams with a modern fish ladder are considered to allow for adequate aquatic animal movement from the impoundments to the wider environment, and typically these would receive $> 70\%$ (Figure 3.8). Fish ladder design and passage has been a controversial issue for some time and information on the effectiveness of fish ladders is a valuable aid in determining this variable [98].

Figure 3.8: Efficient and effective fish pass near Pitlochry in Scotland, which is a minor barrier to aquatic animal passage.

3.3.6 Land Animal Passage (%)

The Land Animal Passage variable is intended to provide information on how easily terrestrial animals such as deer, squirrels and birds can navigate across a

dam or around a water body. This variable requires consideration of the structure of the SFRB and the wider landscape context along with any natural or man-made barriers such as the dam itself. The basin location, dam height and length, fencing, gates, bridges, paths and thickness and type of fringing vegetation can be important factors in this assessment.

Sustainable flood retention basins located in the remote and/or upper reaches of catchments where there is generally sparse population and infrastructure typically represent areas where terrestrial animal passage is good and attract values between 70% and 100%. It is possible for very large dams, often crossing steep valleys, to pose a significant barrier to animal movement due to the high dam structure and the large size of the impoundment it creates. Spillways often create a break in the dam wall, which can be difficult for animals to cross. Sustainable flood retention basins located in urban areas and near roads may pose a significant barrier to terrestrial animal movements [169]; typically, these circumstances result in values between 0 and 20%.

Disused water supply reservoirs are often used for fishing and other recreational activities. Moreover, they may even be designated nature reserves. Such dams often have bridges crossing the spillways and are managed to remain in a natural state, and paths and fencing are put in place to limit human disturbance and increase site safety. Some large natural water bodies can be barriers to animal movement, and in these circumstances a value between 21 and 69 % is typically awarded.

3.3.7 Flood Plain Elevation (%)

Flood plain elevation is defined as the maximum additional height (h_f) that the water rises above the normal height of the basin (h) to reach the flood plain (if present) (Figure 3.9)

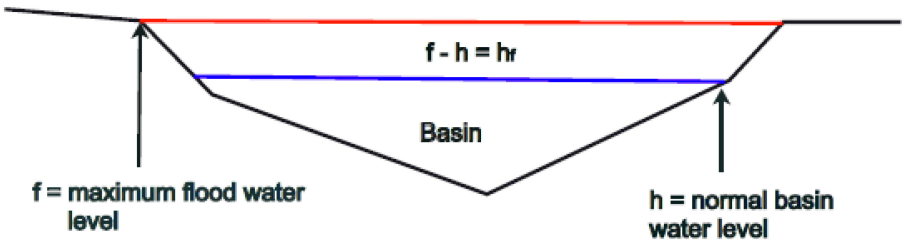
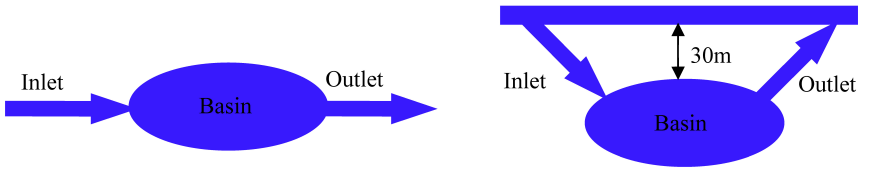


Figure 3.9: Diagrammatic representation of flood plain elevation and flooding depth.

Flood plain elevation is determined during the site visit. It is usually possible to estimate the normal water level of an impoundment or lake by the distribution of debris and water marks around the edges. These can be hidden by high water levels, which can make the variable difficult to estimate. It should, however, be noted that the grass level is often a good indicator of maximum flood plain elevation as this plant cannot tolerate long periods of submergence. A clear line where the grass ends is often a good indicator of maximum flood water level.

In the case of many dams, there is a spillway present. In all such cases the site has a 0 m flood plain elevation as the spillway ultimately sets the impoundments capacity. In some parts of Scotland, such structures have been seen with estimated water depths of between $0.1m$ and $0.3m$, and this observation suggests that a flood plain elevation of $< 0.3m$ is appropriate. Flood plain elevations for other types of SFRB are typically in the range between $0.3m$ and $3m$ based on experience of some water bodies in the west of Scotland. Some of these water bodies flood most winters and achieve additional depths of at least $3m$.



Basin and Channel connectivity of 0 m; i.e. the Basin is on-line with the watercourse

Basin and channel connectivity of 30 m indicating an off-line basin

Figure 3.10: Diagrammatic representation of flood plain elevation and flooding depth.

3.3.8 Basin Channel Connectivity (m)

The basin channel connectivity considers how a SFRB is connected to its water inlets and outlets, and whether it provides a direct path for water flow during flood events. It is an estimate of how directly connected the SFRB is to its water supply and main drainage route, that is whether the basin is on-line or off-line. For an on-line structure, the entire inflow water stream fills and flows through the basin easily, while for an off-line basin, the flood water by-passes the impoundment via an additional channel. The distance between the by-pass channel and the main stream bed flowing through the SFRB is named the basin channel connectivity (Figure 3.10).

An on-line SFRB will have a water inlet and outlet that are virtually part of the river system; i.e. the river effectively flows straight through the SFRB or water body. Such SFRB and natural water bodies receive a score of 0 m.

In the case of some purpose built off-line SFRB, they may only receive water when the river reaches significant flood volumes. These basins are typically built for long return period flood events and are located adjacent to the river. Such basins are considered to be off-line and the distance of this offset is recorded because as offset basins typically have a lower negative ecological impact on the river than those that are built in-line [27, 28].

3.3.9 Wetness (%)

The variable *wetness* has been added to aid in distinguishing between permanently flooded SFRB such as drinking water supply reservoirs and industrial impoundments, and purpose built SFRB, which are predominately dry and are only flooded occasionally in response to major storm events, when they fulfil their flood control function. Many current and former drinking water supply reservoirs effectively run at their maximum design capacity, discharging down their spillways. Such SFRB typically receive a value of >90% depending on how much of the basin they occupy. Deep natural water bodies typically receive similar values. Some shallow SFRB are silting up due to a range of factors including natural landscape processes such as siltation aggravated by eutrophication [116]. In some Scottish locations, this process is accelerated by the removal of arable farming restrictions in the catchment once a drinking water supply reservoir is no longer used for its design purpose. These basins are typically shallow and boggy with extensive fringes of dense reeds and other macrophytes, and they may therefore have very little open water. Such sites typically receive a value between 10% and 74%, depending on the proportion of open water present (Figure 3.3).

Purpose built SFRB designed for long term flood events may only be partly flooded once or twice per decade. Such sites are typically dry basins with high levels of vegetation, and some may be used predominately for other purposes such as recreation or farming (Figure 3.1). These dry basins typically receive a value of between 0 and 9%.

3.3.10 Proportion of Flow within the Channel (%)

This variable describes the proportion of river water that will flow directly through a SFRB, and is linked to the variable *Basin Channel Connectivity* (Section 3.3.8

and Figure 3.10). If an off-line SFRB is present, the mean proportion of flow through the additional channel needs to be considered during the estimation. An off-line SFRB receives a value <100%, while an on-line SFRB always receives a value of 100%. In the case of disused impoundments, these are often found at full capacity with water leaving via the dam spillway, the majority of the flow will be within the spillway channel and typically a score of 100% is awarded. Natural water bodies and wetlands typically have a single main channel or river draining them and these also receive a score of 100%.

3.3.11 Mean Flooding Depth (m)

This is a combined parameter, which is determined by adding the mean additional depth due to an average flood to the mean depth of the basin. If the SFRB is a dry basin, the mean flooding depth is the mean depth of water when the basin is flooded. Many small impoundments, natural water bodies and SFRB are very shallow basins typically with an average depth <1 m. Such SFRB have relatively low flooding depths and are often well vegetated with dense stands of macrophytes e.g. reeds [11]. Typically, the impoundments overall capacity is many times this volume and such SFRB and reservoirs can have depths >20 m, reflected in a very high dam. For such basins and for some natural water bodies, the mean additional depths due to flooding can be $> 3m$ due to steep sides and constrained outlets.

3.3.12 Typical Wetness Duration (d/yr)

Wetness duration is an estimate of the mean number of days during which the basin (not just the stream if present) is wet within a given year. This variable has been added to distinguish between permanently flooded types of SFRB such

as some drinking water reservoirs and industrial impoundments, and purpose built SFRB, which may only be flooded very occasionally, and are therefore predominately dry systems.

Natural water bodies, drinking water supply reservoirs and other forms of large scale impoundments are often permanently wet, typically receiving a value close to 365 d/yr. It is recognized that dams are periodically drained for maintenance and inspection. However, this does not detract from the predominantly permanently wet nature of these structures.

Many SFRB designed for long flood return events may receive a value as lows as 1 d/yr (or even less). Typically, such impoundments are designed to deal with long return period flooding events such as 20, 50 or even 100 years [120].

3.3.13 Estimated Flood Duration (d/yr)

Estimated flood duration is different from the *Typical Wetness Duration* variable (Section 3.3.12), as it only considers the mean number of days in a given year that the SFRB is actually flooded rather than being wet. Typically, this variable is estimated from information based on rainfall patterns and the variable *Seasonal Influence* (Section 3.3.22) to arrive at a probable number. If information is available on water levels for a SFRB, this can be used to accurately determine the number of days of flooding.

In areas of high annual rainfall ($\geq 2m/yr$) and with even a moderate *Seasonal Influence*, there can be relatively frequent flooding events, and sites may be flooded as much as 20 to 30 d/yr (e.g., west coast of Scotland). In areas of low rainfall ($\leq 0.4m/yr$) with a low *Seasonal Influence* such as Mediterranean climates, flooding is likely to have a very low duration of $\leq 2d/yr$.

3.3.14 Basin Bed Gradient (%)

The basin bed gradient is the mean slope of the basin from the main inlet to the main outlet points. Ordinance survey maps should be used to determine the elevation of the basin at each end, and the maps may also be used to determine other variables such as the length of the basin. The basin bed gradient can be calculated by the use of Equation (3.1).

$$g = h_g / l_g \tag{3.1}$$

where g means gradient (-); h_g stands for difference in height between the slope at the head and the foot of the water body (m); and l_g means the length of basin (m).

3.3.15 Mean Basin Flood Velocity (cm/s)

The mean basin flood velocity is defined as the average speed of the water traveling through the entire basin from the inlet to the outlet during a flood. An 'educated guess' is usually used to estimate this value with the support of other variables such as the slope of the basin. Other means of investigation are too expensive and time consuming for a brief investigation.

A high value for this variable is associated with purpose built SFRB located in upland areas where heavy flooding occurs and where a large basin gradient is apparent. The value for such basins can be as high as 150 cm/s. In comparison, a typical value for basins in lowland areas is below 15 cm/s.

3.3.16 Wetted Perimeter (m)

The wetted perimeter is the length of land and solid material that the water in the basin comes into contact with. Such components that are included in the total length are the entire perimeter of the basin, any islands that are within the basin and any vegetation (e.g. tree trunks and reed stems) that are protruding through the surface of the water. The use of ordinance survey maps can help to roughly determine the wetted perimeter of large, deep and geometrically simple water bodies. The higher the wetted perimeter in comparison to the basin area, the higher usually is the diffuse pollution removal capacity.

A brief experiment can be undertaken to estimate the perimeter of reeds; e.g., three small square frames of $10cm \times 10cm$ should be placed around a representative section of reeds to obtain a composite estimate [175]. The wetted perimeter for a very small basin such as a SUDS pond could be below 100 m. In contrast, the wetted perimeter for a much larger basin including islands and vegetation could have a value of over 10000 m.

3.3.17 Maximum Flood Water Volume (m^3)

The maximum flood water volume is reached when a basin is flooded to its maximum capacity and can retain no more water without it spilling over into another basin or catchment. The two main variables that someone should focus on when calculating the *Maximum Flood Water Volume* are the *Mean Flooding Depth* (Section 3.3.11) of the basin and the *Flood Water Surface Area* (Section 3.3.18).

The numerical value for this variable depends predominantly on the size of the water body. If the surface area is small, then the volume will also be small in

comparison to a water body that has a much larger surface area. For upland areas, the water depth of an SFRB is relatively more important than the surface area, while the opposite is the case for lowland areas.

3.3.18 Flood Water Surface Area (m^2)

This is the mean area of the water surface when the basin has been flooded. This information cannot be found on a map as the water surface on the map is often based on the maximum or mean depth. Therefore, an estimate of the flood surface area has to be drawn onto a map and the surface area should subsequently be calculated from this drawing.

Depending on the surrounding landscape, some flood water surface areas can be much larger than the existing mean surface area. This is particularly the case for lowland areas. In contrast, for many upland locations within steep valleys, the flood surface area remains fairly similar to that of the actual surface area of the water.

3.3.19 Mean Annual Rainfall (mm)

The mean annual rainfall [17] is the long term average of the depth of rain that falls within the catchment area within a given year. This information cannot be gathered from a site visit alone. However, the value can be obtained from a database (e.g., Meteorological Office). For the UK, the Flood Estimation Handbook CD-ROM contains exact rainfall data [21, 118].

3.3.20 Drainage (cm/d)

This variable represents how efficiently water moves through the unsaturated zone of the soil and away from the basin. It estimates the mean distance at which water can drain through the unsaturated zone of soil within a typical day.

Drainage can be estimated in a variety of ways with different accuracies. If the soil series around the basin can be identified, the drainage should be characterised from its known drainage properties. Groundwater vulnerability maps are widely available for EU countries, and these can be used to give a general indication of the drainage in an area. Areas where groundwater is considered vulnerable to pollution typically have excellent drainage, and there are few retarding reactions in the overlying soil. Therefore, these areas can be considered to have good drainage. Equally, areas of low groundwater vulnerability are typically those with poor drainage properties or where a layer of clay or other impermeable material underlies the soil [136]. There are exceptions to these general conditions such as cases of extensive and deep organic soils, which protect groundwater from potential pollution or thin soils underpinned by hard igneous rocks. It is important to assess the soil while on site to confirm the initial desk study assessment.

3.3.21 Impermeable Soil Proportion (%)

Permeability is the ease with which a soil allows water to pass through it and is largely determined by the soil type surrounding a basin. Highly permeable strata such as sand and gravel easily allow water to pass through with little retardation. Such highly permeable strata typically contain $< 2\%$ clay. In areas dominated by other soil types, there will be variable and significantly higher levels of clay with

the most impermeable soils being those with a high proportion of clay. Heavy clay soils generally have very poor drainage characteristics [9].

Soil properties can be estimated in the field and there are a range of guidelines to estimate soil types. An easy to use example is the Soil Texture Fact Sheet [18]. The clay content is used in conjunction with the amount of rock present in the soil to arrive at an estimate of the overall impermeable proportion of the soil surrounding an SFRB.

Alternatively, soil principles for classification [26] can be used to determine the soil type and the proportions of sand, silt and clay within that soil. These findings need to be compared to the soil in the field to determine the proportion of rock present.

3.3.22 Seasonal Influence (%)

The variable Seasonal Influence (%) is a sum parameter, which is easy to estimate if general knowledge of regional weather conditions and the landscape topography is available. Climatic conditions have a pronounced effect on whether a basin is wet or dry, and the frequency of flooding. For example, in the UK, winter periods are wetter than those encountered during summer. Most lowland areas around the Mediterranean can be considered to have little seasonal variation and therefore SFRB located in these areas are considered to be subject to low Seasonal Influence (<20%). In contrast, SFRB located in mountainous areas such as the Alps or in regions that are subject to wet climates such as the west coasts of Scotland and Norway, there may be a highly pronounced Seasonal Influence (>70%). Consequently, SFRB identified in these areas are considered to be subject to high seasonal variations [31].

3.3.23 Altitude (m)

This variable is a measure of the altitude at which the site is located and is simply determined from a spot height on a map or by using geographical positioning system equipment. The site elevation for a reservoir with a dam is taken at the bottom of the side of the dam facing upstream.

3.3.24 Vegetation Cover (%)

This variable refers to the basin and not to the corresponding catchment. Dry SFRB can be completely vegetated and, if maintained, are typically grass covered and achieve a vegetation cover value between 20% and 70%. Mature vegetation such as a full basin covered with trees is associated with a value close to 100% (Figure 3.1). In contrast, no vegetation or tarmac will result in values close to zero. It is worth noting whether the SFRB is maintained and vegetated with short grass, which will not impede water flow through the basin, or if the basin is fully covered with mature vegetation, which may reduce its capacity and slow the movement of flood water.

In the case of wet basins, the area of the basin occupied by emergent and floating plants is estimated during the site visit. Sustainable Flood Retention Basins are highly diverse and some may contain values as low as 1% in the case of steep sided maintained drinking water reservoirs to >90% for basins, which are silting up naturally and that are fully covered by mature reed stands.

3.3.25 Algal Cover in Summer (%)

This variable provides an estimate of the degree of phytoplankton growth in a wet SFRB and is a surrogate for the degree of eutrophication in that SFRB. It is easiest to estimate accurately during a summer site visit. Dry SFRB used solely for flood control purposes may have no potential for phytoplankton growth.

Water bodies, which are rich in nutrients, often undergo one or more extensive algal blooms in summer. In countries with nutrient poor waters such as most upland areas of Scotland and Norway, it is unproblematic to estimate the likely potential for pollution related blooms. It can be more challenging for some lowland areas in central Europe where waters are typically higher in nutrient content and support more extensive algal and surface macrophyte communities [4].

An alternative method is the use of a Secchi disk, which is a white plastic disk (15.3 *cm* in diameter) on a weighted line with an accurate depth scale. The disk is simply lowered into the water and the depth at which it is no longer seen is recorded. The more eutrophic the water body, the shallower the depth at which the Secchi disk disappears. The classification used by the Scottish Environment Protection Agency is provided in Table 3.1. However, this approach is of limited value for highly eutrophic waters.

3.3.26 Relative Total Pollution (%)

Pollution is typically defined as the introduction by man of substances or energy that can have a deleterious effect on the environment. In the context of this survey method, the variable relative total pollution is a measure of how impacted a SFRB is by predominantly diffuse agricultural and urban pollution, and it is

Table 3.1: Relationship between Secchi disk depth and eutrophic status for water bodies in Scotland (UK). TS (Trophic Status); MTP (Mean Total P ($\mu g/l$)); MT (Maximum Total ($Chl_a(\mu g/l)$)); MSD (Mean Secchi Depth (m)); MiSD (Minimum Secchi depth (m)).

	TS	MTP	MT	MSD	MiSD
Oligotrophic	< 10	< 2.5	< 8	> 6	> 3
Mesotrophic	$10 - 35$	$2.5 - 8$	$8 - 25$	$3 - 6$	$1.5 - 3$
Eutrophic	> 35	> 8	> 25	< 3	< 1.5

largely a function of the breakdown of land types within a catchment and the way that the land is used.

Diffuse agricultural pollution is a significant problem in many areas of Europe. The pollution arises as a consequence of the normal arable and livestock application of fertilisers and agrochemicals as management tools [173]. Some of these chemicals are transported by surface and groundwater movement and can be captured within SFRB, particularly those designed to trap sediment. The degree of runoff and impact depends on farming practices, chemical application rates and tillage practices. Livestock farming can result in significant inputs of nitrates and phosphates from the animals, and corresponding microbiological contamination can be a problem during storm events [36].

Old mines and associated spoil heaps and processing areas can be significant sources of pollution of the water environment. There is a substantial history of mining for most minerals and metals throughout the world. It is therefore valuable to consider whether there may be any water contamination from this source, which may affect the pollution status of an SFRB [169]. The presence of a mine near an SFRB could result in a pollution assessment from anything from 2% for an old mine with no visible impacts through to 100% for a site contaminated

by acid mine drainage. Industrial processes can be a source of pollution within catchments. In Europe and North America such facilities are closely regulated and monitored so that they do not exceed strict consent conditions [128]. These facilities would be included in the assessment and a low value of between 3 and 5% pollution would be associated with such a site. Old, derelict industrial facilities can be associated with significant land contamination problems, and such sites should therefore be considered to have a high pollution potential [111].

Many SFRB including larger SUDS impoundments such as wetlands and retention basins can be found in urban areas. These SFRB typically have a tacit diffuse pollution control and mitigation function as they are designed to receive the first foul flush of contaminants associated with storm runoff. Typically, these contaminants such as poly-aromatic hydrocarbons (PAH), mineral oil and grease from road traffic are trapped in the basin, and once the basin is dry, these can be broken down by photolysis. Contaminants such as metals tend to build up in the sediments of these basins, and these can pose a problem for waste disposal in highly contaminated areas [158].

The overall level of pollution of a basin is assessed by taking all of the factors outlined above into account. The site visit is particularly valuable in this regard as it shows land use practices and often reveals the industries present within a catchment.

3.3.27 Mean Sediment Depth (cm)

The mean sediment depth is the average depth of the sediment within a SFRB structure whether wet or dry. The sediment depth can simply be determined in dry basins by digging a shallow hole and subsequently assessing the sediment profile.

In permanently flooded SFRB, the sediment is not accessible, and an estimate of the sediment depth is therefore derived. Freshwater sedimentation rates are highly variable and depend on a wide variety of factors. For example, research has revealed that sedimentation rates for oligotrophic waters are as low as 0.16 mm per year, though these can increase to a range between 5.6 and 11 mm per year, when pollution by phosphate and nitrate occurs [122].

In the absence of published information, an estimate of the sediment depth can be made using expert judgement based on the sedimentation rates outlined above (depending on eutrophic status) multiplied by the number of years that the SFRB has been present. It should be noted that water supply reservoirs are generally located in areas where there are low maintenance costs associated with siltation and are managed to maintain the vegetation cover within a catchment.

It is standard practice in limnological investigations to express sedimentation rates as grams of sediment per meter per year. If the bulk density of the sediment is known, it is a simple matter to convert this to the depth of sediment [137]. The research team is currently actively working on the development of a better method for sediment depth estimation based on simple field measurement techniques.

3.3.28 Organic Sediment Proportion (%)

The proportion of organic matter within sediment is determined by complex interactions in the water environment of a catchment. Organic material is provided by terrestrial plants and animals, combined with the production from aquatic algae, plants and animals within a wet SFRB to establish the overall production of the water body. This organic input is then metabolised by bacteria and sediment-dwelling invertebrates, which utilise a large proportion of the organic carbon as an energy source. The final organic proportion of the sediment

is an interaction of these metabolic processes and the deposition of gravel, sand, silt and clay from within the catchment [96].

In upland areas with high rainfall, there are usually relatively low proportions of organic matter present in the sediment. This type of upland catchment tends to host oligotrophic water bodies with limited primary and secondary production, and the majority of the organic matter within the basin is cycled through the biota [119]. These types of catchments tend therefore to have low proportions of organic matter present within the sediment and would typically receive organic sediment proportion values between 1% and 3%. An exception to this general rule can be where dense conifer plantations have acidified the water in an upland catchment. In these cases, refractory conifer leaf litter is a feature of all visible sediments [7], and such sites should be considered to have a very high organic sediment proportion of >7%.

In lowland areas and for areas where the waters are oligotrophic to mesotrophic, sediments are considered to have relatively high organic matter content due to the increased productivity of such waters, and these are typically assigned values between 7% and 15%. Where waters are eutrophic and highly productive there is often significant deposition of organic matter at the bottom of a permanently wet SFRB [23]. These sites are considered to have a relatively high organic matter content of between 16% and 30%.

Hyper-eutrophic water bodies are those which are permanently polluted by sources of nutrients such as sewage discharges or from diffuse or point source pollution. These water bodies have a high primary production and regularly suffer from algal blooms. When dead algae sink to the bottom of a water body, they tend to cause anoxic conditions, reducing the breakdown rate of organic materials in the sediments. Concurrently pollutants such as nitrate, phosphate

and ammonia are released into the water body. In such situations it is possible to find organic matter at between 30% and 60% of the sediment [196].

Values >60% of organic sediment should not routinely be assigned to SFRB unless there is clear and compelling evidence that this is the case. For example, an exception is where a SFRB is surrounded by peat bogs. Peat bogs are decayed and water logged sphagnum moss and other plant materials that have partially decomposed. This type of soil contains virtually 100% organic material [182].

The ideal solution to determine the organic carbon content of sediment is by direct measurement of a homogenised sample, typically achieved using an organic carbon analyser, which heats the sample and converts the organic matter present to carbon dioxide that is then measured [164]. This measurement could be used directly in the SFRB classification system.

Due to the complexities of sediment dynamics and particular geo-chemical circumstances, the above boundary conditions may not be fully applicable to the survey area of interest. It is recommended that the survey team considers whether the boundaries proposed above are suitable for their survey area. Information on areas with sedimentation problems can be found in river basin management plans prepared by the European environment agencies [33].

3.3.29 Flotsam Cover (%)

Flotsam is defined as debris and waste that is floating on the surface of a water body. It can include items such as debris from tress, rubbish thrown into the water by humans or even abandoned boats or submerged cars. The principle objective of this variable is to determine if there is flotsam present that might restrict the flow of water out of a SFRB.

Levels of flotsam vary widely from upland catchments where there may be virtually no flotsam present through to urbanised SFRB that have been used as dumping grounds for cars, shopping trolleys or other man-made debris. Ultimately, the value awarded for flotsam cover should reflect the proportion of cover of the outlet by flotsam. This variable is an indirect measure of flow restriction. Therefore, basins with outlets that have little flotsam cover receive low values (typically <5%), where basins, which are full of flotsam such as some invasive plant species or a lot of human rubbish would receive high values, if the outlet structure is severely covered with flotsam.

An important aspect to verify during a site visit is whether there are screens present on overflows to retain fish, and whether these are clear of debris or obviously clogged. In cases where a fine mesh type screen is in place to retain fish these can become clogged with leaves or items such as plastic bags. Clogged screens can raise water levels within a basin taking it above its design capacity.

3.3.30 Catchment Size (km^2)

The catchment size is the area of land from which water feeds into an SFRB. Information on the catchments that feed individual dams can be found on national registers of these structures and if available should be used as a high quality data source. In the absence of a national register of dams and reservoirs, the European environment agencies have defined river basin districts and sub-catchments within the larger units [47]. These are a potentially valuable resource in the assessment procedure.

If access to a geographical information system of the survey area is available, the system can be used to define the likely catchment area for the SFRB based on the topography of the base map. The simplest method is to use commercially

available digital map packages. Typically, these have the ability to translate the map into a three-dimensional terrain model. The area of a catchment can then be marked on the map by using the hills and streams to define a likely catchment area. This can be somewhat more problematic for lowland areas where there is a predominately flat topography. Paper maps can be used in the same way, and the area is then directly estimated from the map.

3.3.31 Urban Catchment Proportion (%)

Water quality within any catchment is the result of a complex series of interactions between the water source, water properties, catchment geochemistry and inputs from human activities. As this is a rapid screening method, a range of land types, which are considered to have different polluting properties, have been defined for the temperate survey areas. In some parts of the world, these land use categories are not appropriate and different land use types may need to be included. The most important feature of such substitutions is that the land type has a known runoff and pollution potential and is widely applicable within the survey area.

The urbanised proportion of a catchment, for example, is simply the area occupied by man-made structures such as roads, farms and towns. The urban catchment is likely to be an important source of diffuse pollutants to the water environment and is found globally.

In the case of SFRB, which have a very high urban catchment proportion (>90%), it may be appropriate to replace the natural land use categories below with different types of urban development. Such a division could include light or heavy industry, industrial estates, retail and residential areas as appropriate.

3.3.32 Arable Catchment Proportion (%)

Arable land is defined as areas where crops are grown either for commercial agricultural purposes or subsistence farming. The type of crop is likely to vary with climatic conditions and weather and can be used for any type of farming where rows of crops are interspersed with bare soil. In some parts of the world, a separate category for rice farming and aquaculture will be needed.

A high proportion of arable area generally contributes greatly to the diffuse pollution of water bodies, so the *Relative Total Pollution* (Section 3.3.26) of the basin and the *Organic Sediment Proportion* (Section 3.3.28) might be high.

3.3.33 Pasture Catchment Proportion (%)

Pasture is land where animals are taken to graze. In temperate climatic zones such as Northern Europe, pasture land is typically managed and consists of relatively short and dense grass which looks like a monoculture. It is given a separate category as it can be a significant source of nutrient and microbiological contamination of the water environment.

3.3.34 Viniculture Catchment Proportion (%)

Viniculture is the practice of growing vines to produce grapes, which are later made into wine. Many vineyards are on steeply sloping land and have relatively bare soils. These conditions can result in significant runoff and soil erosion if badly managed [20]. In many areas of the world, viniculture is a common practice; however, this variable is not relevant to cool temperate climates such as Scotland and Norway.

3.3.35 Forest Catchment Proportion (%)

This variable is simply the proportion of the catchment that is covered by predominantly managed forest, can easily be estimated or measured using maps and should be ground-truthed during a site visit. In heavily forested catchments, it may be desirable to distinguish between natural woodlands and forestry plantations, particularly where forestry plantations can be a significant source of diffuse pollution and acidification [124].

3.3.36 Natural Catchment Proportion (%)

This final category of land use is intended to cover the remaining proportion of the catchment and is considered to be the land where there is no or minimal interference from man. It may therefore constitute potentially remote grassland, scrubland, moor and similar types of land.

Generally, basins located in the upper lands or in the deep valleys have high *Natural Catchment Proportions*. In contrast, basins located in urban areas always have very little or no natural features. This variable can be assigned a value during the site visit when comparing its characteristics with those of the corresponding rather urban and forest catchment proportions, all of which can be obtained from maps. A high fraction of natural catchment also correlates positively to a low *Organic Sediment Proportion* (see Section 3.3.28).

3.3.37 Groundwater Infiltration (%)

Groundwater is considered to be the water that lies beneath the saturated zone of the soil and is composed principally of surface waters and rainfall that has

percolated through the soil and into the underlying rocks and typically intersects with water bodies such as SFRB and natural lakes [110]. This parameter indicates the proportion of the water within an SFRB that comes from groundwater, and it can be a significant source of water for some cases outlined below.

In the case of purpose built dry SFRB, there is virtually no groundwater infiltration. Moreover, former industrial impoundments and drinking water supply reservoirs are typically lined with an impermeable layer of clay, and are therefore isolated from the surrounding groundwater. Both types of SFRB receive a value of $< 5\%$.

Some wet SFRB and natural lakes may receive a fair proportion of their water from groundwater. Such basins are typically shallow and due to the groundwater flow contain a small lake or pond within a larger basin. This is generally most apparent during the drier period of the year where springs can become visible. Such basins typically receive a value of between 5% and 10%, with 10% being considered a typical value for a natural lake or pond.

Basins may be encountered in some regions where groundwater composes between 10% and 40% of the flow of water from the SFRB, and these are considered to have a high dependence on groundwater. Very high values for groundwater infiltration to a basin would be between 41% and 50%.

There may be special cases where SFRB are encountered which receive >50% of their water supply from groundwater. These systems have usually no stream as an inflow source. At the time of writing, only one such example is found where a Scottish SFRB is being kept permanently wet by receiving groundwater. This basin is located in an area where there is significant hexavalent chromium contamination. Hexavalent chromium is a human carcinogen by inhalation and therefore dust arising from a dry SFRB could pose a hazard to human health.

Utilising the groundwater to keep the basin wet is a novel approach which minimises this risk. Approximately 80% of the corresponding SFRB water budget comes from groundwater, with the other 20% being supplied by road drainage and direct rainfall.

3.3.38 Mean Depth of Basin (m)

The mean depth of the basin is simply the average depth of the impoundment. In the case of dry SFRB, the depth should be recorded as the possible mean depth to aid in computing the flood water capacity of the basin.

In the case of permanently flooded SFRB, the corresponding depth has to be estimated. Excellent sources of information on water depths are fishermen who regularly fish the SFRB. Anglers are a passionate and knowledgeable group, and if a fishing club has a lease on a SFRB, its members generally have extensive understanding of its history as well as physical characteristics, which can be invaluable to the surveying team. For very large SFRB, there may be accessible sources of survey information to provide the average depth of a water body. Where this information is available, it often contains information such as maximum depth. In the vast majority of wet SFRB, the mean depth of the basin will need to be estimated.

3.3.39 Length of Basin (m)

The length of the basin is the distance from the two points of the basin perimeter that are furthest away from each other at normal environmental conditions (e.g. no flooding). In an ideal case, this is the distance between the inlet and the outlet.

3.3.40 Width of Basin (m)

The width of the basin is the distance across the basin at normal environmental conditions, and is ideally at right angles to the basin length (see also Section 3.3.40).

3.3.41 Dam Condition (%)

The variable Dam Condition is intended to be empirical, largely based on site visits and photographic evidence. The assessor should record the overall dam condition and associated maintenance undertaken to maintain the structure. The higher the score, the better will be the dam condition. A composite score will be based on the following components associated with different weightings:

Dam Structure (%) accounting for 30 percentage points overall. This variable is intended to assess the overall condition of the dam. Account should be taken of the dam size, material used (e.g. concrete, rock and earth) and how tidy the dam appears as a surrogate for overall maintenance. The face of the dam should also be examined for stability and any obvious signs of surface cracking. If surface cracking or seepages are apparent, then a low score should be awarded.

Spillway Condition (%) accounting for 30 percentage points overall. Spillway failures can be a significant cause of overall dam failure should water penetrate and begin to erode the dam face. Spillways are typically masonry or concrete structures, though in some cases, hybrid structures can be found. Concrete spillways should typically receive a higher score than masonry structures, unless they are poorly maintained. Masonry spillways are, however, safe as long as properly maintained. Spillways that are obviously poorly maintained typically have a high proportion of vegetation growing through the pointing. In cases

where masonry dams are obviously missing blocks, a score of between 0 and 5 should be awarded.

Wave Wall Condition (%) accounting for 20 percentage points overall. The wave wall is an essential component of earth dams and is typically a masonry wall lining the front face of the dam. The masonry is essential in preventing erosion of the earth dam itself. Therefore, its condition is vital for the safety of the dam. A well maintained wave wall with no visible vegetation would receive a score of close to 100. The proportion of pointing (i.e. mortar or cement between stones) containing vegetation should be taken into account when estimating the overall score. In the case of a concrete dam, the inside face of the reservoir should be assessed and its general visible condition determined.

Operational Volume Impact (%) accounting for 10 percentage points overall. Dams are not designed to be continually maintained at their maximum volume indicated by spillways continually discharging. Dams maintained in this condition should be assigned a value close to 0. For an initial assessment, the previously determined variable *Maximum Flood Water Volume* (see Section 3.3.17) could be standardized (after removal of outliers) and the corresponding values may be taken times 100 to obtain the *Operational Volume Impact*.

Other Factors Influencing Dam Condition (%) accounting for 10 percentage points overall. Other site-specific observations not related to the points mentioned above may receive an overall score of not more than 10. For example, the overall site management and maintenance is often a good indicator of dam condition. Furthermore, off-line reservoirs are likely to contribute less to flood risk than on-line reservoirs, which have been constructed on the original river bed. Moreover, reservoirs currently operated by Scottish Water for drinking water purposes are likely to score around 90%, which reflects Scottish Water's commitment to the

safe operation of their water infrastructure. Other operators such as councils, fishing clubs and sailing clubs may have less stringent safety standards.

3.3.42 Dam Failure Hazard (%)

The *Dam Failure Hazard* variable is intended to provide an overall estimate of the potential damage resulting from dam failure. High scores are assigned to dams where the risk of failure is likely to be high. Relevant information for this variable can be obtained via site visits and careful assessments of the geographical characteristics downstream of the structure in combination with the outputs from hydrological and hydraulic models [166]. The damage potential is particularly affected by hazardous processes, particularly if they vary spatially and/or temporally. A site posing the maximum hazard would be a large dam of more than 25,000 m^3 located within or just upstream of a dense urban area with housing immediately down gradient of the dam or reservoir. Sites such as the drinking water storage reservoirs in Milngavie (near Glasgow, Scotland) would receive a very high hazard score of close to 100. This approach is consistent with the A-E reservoir hazard ratings used in the Reservoirs Act 1975 (as Amended) as outlined by the Office of Public Sector Information [19]. In contrast, a small dam located in an upland area, where dam failure would follow a river valley (potentially without affecting any housing or other infrastructure), would receive a very low score in the range of between 0 and 10. The following components and associated weightings for *Dam Failure Hazard* have been proposed:

Overall Force on Dam (%) accounting for 30 percentage points overall. The higher the overall force on the dam, the more likely will be a dam failure. The *Overall Force on Dam* component is largely affected by the previously defined SFRB variables *Dam Height* and *Dam Length*.

Potential Loss of Life (%) accounting for 35 percentage points overall. Essentially, the more people would be affected by flooding, the higher would be the score for this component. An indication for potentially high losses would be urban areas in the catchment just below of the failed dam. A useful source of information to estimate *Potential Loss of Life* is, for example, the Indicative River and Coastal Flood Map for Scotland [166]. Moreover, Haynes et al. [68] predicted the social impact of flooding by using statistical evaluation techniques of census data.

Importance of Infrastructure Affected by Dam Failure (%) accounting for 25 percentage points overall. This component scores high if important infrastructure would be affected by flooding due to dam failure. Infrastructure elements may comprise airports, railways, major roads, retail parks, farming infrastructure and assets, and water and electricity supply structures. A useful source of information to estimate *Importance of Infrastructure Affected by Dam Failure* is the Indicative River and Coastal Flood Map for Scotland [166] in combination with detailed geographical maps.

Other Factors Influencing Failure Hazard (%) accounting for 10 percentage points overall. These factors are very much site-specific, and may include unusually poor dam conditions such as erosion and damage due to rodents; e.g. "honeycombing" of embankments as a result of rabbit burrowing [52]. Moreover, nature reserves protecting endangered species and amenity areas could get destroyed. Note that it is important to distinguish between hazard and risk. The hazard associated with a dam considers only the consequences of the structure failing. The risk of that failure (hazard) occurring is determined by factors such as the management of the dam and its maintenance, potentially increasing the probability of failure. However, no exact equation can be used to accurately determine this complex variable, and expert judgement is therefore needed.

3.3.43 Dam Failure Risk (%)

In the area of natural hazards, risk is defined as a function of probability of occurrence, intensity, extent of damage and vulnerability. Furthermore, risk assessments also take geographical and statistical data on elements at risk into account [171].

The European Flood Directive [39] recommends the creation of flood risk maps with different hazard criteria: (a) flood events with a high probability (HQ 10); (b) flood events with a medium probability (HQ 100); and (c) flood events with a low probability (extreme event). Any risk variable should also include information on water depth and velocity [42] as well as consideration for areas with embankment erosion and sediment deposition.

The number of potentially affected inhabitants and business activities as well as environmental damage should also be depicted. Hence, vulnerability indicators may vary between low (such as agricultural areas and individual farm estates), medium (such as dispersed settlements and small villages), and high values (city centres and industrial zones) as discussed by Spachinger et al. [171]. Moreover, economical scores differentiating between industries under varying scenarios could be used to inform *Dam Failure Risk*. For example, a paper mill storing chemicals is of greater importance than a garage.

The variable*Dam Failure Risk* is intended to capture the risk of a major structural failure. Therefore, this variable has to consider the hazard posed by the structure, and how it is maintained and managed. Utilising this approach, a composite variable can be derived based predominantly on the variables *Dam Conditi*on and *Dam Failure Hazard*. However, no accurate single equation can be used to determine this complex variable accurately, and expert judgement on various risk

components is therefore required. The following components for *Dam Failure Risk* have been proposed:

Failure Risk (%) accounting for 20 percentage points overall. The more neglected a dam is due to poor maintenance, the more likely will be a dam failure due to too high pressure on the dam during a flood event of considerable high duration. A well maintained dam with appropriate materials used and a safe operational mode would receive a low score, and a poorly maintained unsafe structure a high score.

Loss of Life Risk (%) accounting for 50 percentage points overall. Essentially, the more people who would actually be present in an affected area during the flooding event, the higher would be the score for this component. Furthermore, the flooding depth and water velocity would need to be considerably high and rapid, respectively, over long periods [42]. An indication for potentially high losses would be dense urban areas with a high proportion of permanent population in the catchment just below of the failed dam [171]. For an initial relative indication of *Loss of Life Risk*, someone may wish to multiply *Failure Risk* times *Potential Loss of Life*. However, the result would need to be revised considering probabilities for various dynamic scenarios.

Risk of Infrastructure Failure (%) accounting for 20 percentage points overall. This component scores high if important infrastructure that is poorly protected against flooding would be affected by dam failure. Furthermore, the flooding depth would need to be considerably high over long periods. Infrastructure elements particularly at risk comprise airports with low-lying runways close to watercourses, railways with low embankments, major roads with bridges through valleys, and water and electricity supply structures in lowlands that are close to watercourses. For an initial relative indication of *Risk of Infrastructure Failure*, an assessor may wish to multiply *Dam Failure Risk* with the *Importance of*

Infrastructure Affected by Dam Failure. Nevertheless, the outcome would need to be adjusted subject to the likelihoods of various scenarios.

Other Factors Influencing Failure Risk (%) accounting for 10 percentage points overall. These factors may include excessive embankment erosion during particularly wet years, contaminated sediment deposition in populated areas due to prolonged flooding and damage due to uncontrolled rodent population expansion due to ideal breading conditions [52]. Other factors may also include the risk subject to unforeseen circumstances such as extreme shifts in weather patterns due to climate change, war damage or terror attacks. Moreover, if the failure of a particular reservoir would result in the likely failure of a further reservoir located downstream, a high score should be awarded for *Other Factors Influencing Failure Risk*.

The proposed *Dam Failure Risk* variable has its limitations. Previous research into the spatial and temporal risk of flooding has largely been restricted to empirical estimates of risk measures. For example, Baggaley et al. [8] indicated that an analysis of long-term seasonal data suggested a shift towards increased flows in spring (March to May) and decreased flows in summer (June to August) for the River Dee in Scotland. A weakness with such an empirical approach to risk is that there is no basis for extrapolation of estimates to rarer events, which is often required as empirical evidence suggests that larger storm events tend to be more localized in space. Therefore, Keef et al. [88] adopted a model-based approach, which accounts for missing values. However, as the complexity of flooding risk increases, the number of missing data or even missing variables increases rapidly as well, justifying expert judgment to be made at reasonable expense.

3.4 Assessment of the Primary Purposes of SFRB

3.4.1 Overview

The primary purposes section is intended to establish the roles that the SFRB was first designed for as well as its current uses. It should be clear that not all SFRB are used for their original purpose, for example, in Scotland there are many former industrial impoundments and disused drinking water supply reservoirs. These structures are typically somewhat neglected, which then results in a high biodiversity making them attractive to the local communities who now use them as nature reserves or for other recreational activities such as walking, bird watching and fishing. It is these evolutions and changes in function with time that are captured by this section of the survey method.

3.4.2 Dominant Hydraulic Purposes

A SFRB with a dominant hydraulic purpose is typically either wet or dry. Wet SFRB are purpose built water storage reservoirs for drinking water or for the regular supply of water to industrial processes. Dry types of SFRB with a predominately hydraulic function are those built for long return period flood events, which may be part of an integrated system of flood defence reservoirs as is seen in Baden, Germany [158].

3.4.3 Drinking Water Supply

In the UK, there is an extensive network of current and former drinking water supply reservoirs and these are a significant landscape feature in some areas. This reflects Scotland's reliance on surface waters to supply drinking water. In many other parts of the world, groundwater is a significant source of drinking water, and therefore surface water reservoirs may be relatively rare. In Germany, for example, approximately 80% of drinking water is supplied by groundwater, while 70% of UK drinking water supplies come from surface waters [29].

3.4.4 Production Industry and Economic Use

Many industries from iron and steel production to chemicals to food stuff manufacturers require high volumes of water to maintain industrial production. Historically, many plants built their own water supply reservoirs and in some cases these impoundments still exist many years after the industry they served has gone. It can be difficult to accurately identify such sites without reference to historic maps.

In the UK, the Ordnance Survey has made all their old maps available digitally and these are a valuable source of information and freely available to researchers through the EDINA university system [35]. Industrial impoundments can be found by comparing the series of historic maps for an area of interest. If similar high quality map data is available for other countries, it should be used as a definitive source of information. Regardless of how the information is obtained, any impoundments built for industrial use would receive a high percentage for this category [163].

Many SFRB in Southern Baden are used for farm land, vineyard or forest, since

they were almost empty and dry throughout the whole year. People use these SFRB group plants, grapes or fruits trees or forestry and then gain incoming from the harvest of food, fruits, viniculture and wood.

3.4.5 Sustainable Drainage

A SFRB may be a large SUDS, where the purpose is to uphold best management practice in terms of infiltration, water quality improvement and sustainable resource management [191], which is the fundamental basis of current thinking on sustainability and seen as best management practice. New SFRB are designed to be sustainable. However, this may not have been a consideration with historic SFRB, which were built before the modern practice of sustainability evolved. It is, however, important to recognise that these older structures may now contribute to sustainable drainage either through hosting a wetland system or by providing retention capacity or if silting up by retaining sediments, which can be regarded as sustainable drainage and a low value of between 0 and 20% can be assigned for older structures.

3.4.6 Environmental Protection

Some water bodies including SFRB may have environmental benefits, and are even protected because they are part of nature reserves or SSSI. These SFRB are mainly for the protection of animals, vegetation and ecology. Normally, these will have signs indicating that there are protected species in the area or have warning signs indicating so (Figure 3.11).

In such cases the score assigned to environmental protection can be as high as 50%,

Figure 3.11: Example of how an old hydraulic retention basin can evolve into a nationally important nature reserve.

with the other 50% of the purpose assigned to Drinking Water Supply (Section 3.4.3) as this was the original design purpose of the SFRB.

Figure 3.12: Example of recreational facilities at a disused water supply reservoir.

3.4.7 Recreational Benefits

The main purpose of some water bodies may be that they benefit the general public by providing recreational benefits including sports, walking, fishing and bird watching. These SFRB can be easily identified as they are often managed; e.g. fishing areas are provided with huts, which provide fishing supplies and check permits of anglers (Figure 3.12).

3.4.8 Landscape Aesthetics

Most natural water bodies and some SFRB, often located in parks, contribute significantly to the aesthetic value of a landscape. These watercourses are looked after by a local authority to maintain the aesthetically pleasing view for walkers and other user groups.

3.5 Summary

This chapter presented the guidance manual of the rapid assessment methodology for survey of Sustainable Flood Retention Basins (SFRB). In this guidance manual, each variable and purpose of the SFRB on the survey template was described and determination of appropriate values for specific SFRB was explained. It provided useful guidance for new users to investigate potential SFRB by using the survey template. In addition, the guidance manual worked as a standard, playing a key role in keeping the assessment of SFRB in consistent and comparable. Furthermore, the assessment approach provided a rapid screening tool to identify water bodies and flood defence structures which have the potential to be used as part of a sustainable flood risk management strategy.

Chapter 4

METHODOLOGY

According to the basic definitions of SFRB (Chapter 2) and the guidance manual to survey SFRB (Chapter 3), data sets of SFRB have been established. To reveal the patterns underlying the collected data, various machine learning algorithms and geo-statistics are explored and applied to SFRB data sets. This chapter summarizes the methodology adopted in the thesis.

The following sections are organized as follows. Section 4.1 starts with an introduction of cluster analysis. Afterwards, principal component analysis, self-organizing maps and feature selection are presented separately in Sections 4.2 to 4.4. Multi-label classification techniques are discussed in Sections 4.5. The spatial statistics, which mainly are ordinary kriging and disjunctive kriging, are demonstrated in Section 4.6. Section 4.7 presents the dam failure assessment tool for SFRB. Finally, this chapter is ended with a summary in Section 4.8.

4.1 Cluster Analysis

4.1.1 Rational

In real life, most natural and constructed retention basins retain runoff for subsequent release, thus reducing downstream flooding problems. However, some basins such as wetlands perform other tangible, albeit less 'visible' roles, including diffuse pollution control and infiltration of treated runoff or promoting groundwater recharge. It means different basins often perform different functions. Moreover, some basins may even have multiple functions simultaneously. From a taxonomic point of view, the functions of retention basins largely determine the types of SFRB. Thus, it implies that the types of retention basins are diverse and rather complicated. The ambiguity of SFRB types may lead to confusion and even conflicts between stakeholders, authorities, engineers as well as decision-makers. Moreover, without expressly learning the real functions and status of SFRB, some significant functions of SFRB (e.g. flood reduction, environment protection) may never be discovered and developed. Therefore, it is highly urgent to distinguish SFRB types and thus to better understand and manage the status and functions of SFRB. To achieve this target, clustering provides an appropriate way of finding the intrinsic group patterns of SFRB data.

Clustering is a widely used method of unsupervised learning for data analysis. It is the organization of a collection of unlabelled data (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity [77]. In other words, the collection of objects in one cluster are similar to each other while are dissimilar to the objects belonging to a different cluster. Specifically, the SFRB that have similar characteristics would be grouped in one cluster whereas the SFRB in different clusters are dissimilar. The intrinsic

clusters of SFRB correspond to SFRB types. Therefore, the types of SFRB can be identified with the help of clustering.

4.1.2 Related Work

Clustering has a rich history in a wide variety of disciplines [78] ranging from biology, psychology, archaeology, geology, geography to marketing. Increasing interest has arisen in many other fields, including machine learning [76], data mining [12], image processing [51], and pattern recognition [77]. Very good references for data mining clustering techniques can also be found in [64, 65].

During the past several decades, many algorithms have been proposed for clustering, including K-Means [112], EM [146], DBSCAN [41], OPTICS [117] and CURE [57]. K-Means and EM are the most fundamental techniques in PDF-based clustering. The basic idea of these methods is to detect clusters by using a model of probability density functions (PDFs) to describe the data structure. However, these methods are restricted to detect spherically or Gaussian clusters and are sensitive to noise and outliers. Moreover, the number of clusters K needs to be specified by the user, which is a non-trivial task on the real-world data sets. Density-based clustering (e.g. DBSCAN and OPTICS) overcomes these problems. The basic idea of DBSCAN is that for each object of a cluster, the neighborhood of a given radius ε has to contain at least a minimum number of objects($MinPts$), where ε and $MinPts$ are input parameters. Although DBSACN can effectively detect arbitrarily shaped clusters, it is difficult to choose a suitable settings for the two parameters, especially since the parameters are correlated. The OPTICS algorithm analyzes the hierarchical data from the perspective of density. It creates the reachability plot to visualize clusters of different densities of the data sets. However, for many real data sets, the reachability plot is very smooth and cannot find the hierarchal clusters. The technique CURE utilizes

multiple representative points to evaluate the distance between clusters to exploit arbitrary shaped clusters. However, for a given data set, it is still difficult to define appropriate splitting levels which correspond to meaningful clusters. Hierarchical clustering (for a detailed survey see [64]) creates a cluster hierarchy, also known as dendrogram. The dendrogram can be formed in two ways: top-down or bottom-up. Accordingly, hierarchical clustering generally falls into two categories: divisive (top-down) approaches and agglomerative (bottom-up) approaches. The divisive clustering starts by viewing all objects in the same cluster. The cluster is split into smaller clusters recursively until eventually each object is in one cluster. The agglomerative clustering works in a converse way. It starts with the clustering by placing every object in a unique cluster, in every step two closest clusters are merged until all objects are in a whole cluster.

Scholz and Sadowski [163] applied Ward's link clustering on 141 SFRB in Baden, forcing them into 6 clusters. This application was found to be limited as it did not allow the clustering to produce the intrinsic groups for the SFRB data. Moreover, the generated 6 clusters did not always match with the 6 predefined SFRB types.

4.1.3 Agglomerative Hierarchical Clustering

Among the clustering approaches, the agglomerative hierarchical clustering has wide popularity. It is simple, straightforward and is firmly based on the foundation analysis of data, thus it was chosen to analyze the SFRB data set in this thesis. To determine which clusters should be merged, it requires a measure of similarity between sets of observations and the linkage criterion which is a function of the pair-wise distances of observations. The similarity (closeness) of the clusters is determined by the measures of distance (e.g. Euclidean distance) between pairs of observations. For the merging criterion, several alternatives have been proposed, the most well known of which are Single Link, Average Link,

Complete Link and Ward's Link. Various partitioning clusterings can be obtained by cutting the hierarchy at various desired levels. The higher the level, the coarser the clustering is, and the smaller is the number of clusters.

Generally, agglomerative clustering can be further briefly classified into 4 types depending on the linkage criteria used. Each type is described as follows.

Single Link: Single linkage clustering is one of the simplest, often used agglomerative hierarchical clustering methods. Suppose two sets of observations (clusters) A and B, their distance is defined as the distance between the closest pair of objects, where one object is taken from A and one from B.

$$D(A, B) = min\{d(i, j)\} i \in A, j \in B \tag{4.1}$$

here object i is in cluster A and object j is in cluster B; $D(A, B)$ means the distance between clusters A and B; $d(i, j)$ stands for the distances between every possible object pair (i, j).

In single linkage clustering, the distance between two clusters is given by the value of the shortest link between any two objects in the two clusters. The height at which two clusters are merged in the dendrogram reflects the distance of the two clusters. At each stage of agglomerative hierarchical clustering, the clusters, whose distance $D(A, B)$ is minimum, are merged to form larger clusters, and so on.

Complete Link: In contrast to single link clustering, the complete link clustering uses a different measure of defining distances between clusters. It defines the distance between two given clusters A and B as the distance of the most distant pair of objects (one from each cluster).

$$D(A, B) = max\{d(i, j)\} i \in A, j \in B \tag{4.2}$$

where object i is in cluster A and object j is in cluster B; $D(A, B)$ means the distance between clusters A and B; $d(i, j)$ stands for the distances between every possible object pair (i, j).

In this case, the maximum value of any pair of objects $d(i, j)$ is regarded to be the distance between clusters A and B. After that, using the hierarchical clustering, the clusters whose distance $D(A, B)$ is a minimum, combine to form larger clusters, and so on.

Average Link: Average link clustering is one of the variants of the above two clustering methods. The mean distance between two clusters A and B is defined as follows:

$$D(A, B) = \frac{1}{|A| \cdot |B|} \sum_{i \in A} \sum_{j \in B} d(i, j) \tag{4.3}$$

where object i is in cluster A and object j is in cluster B; $D(A, B)$ means the distance between clusters A and B; $d(i, j)$ stands for the distances between every possible object pair (i, j), $|A|$ and $|B|$ are the number of objects in clusters A and B respectively.

Here, the distance between clusters A and B is the average value of the distance between any pair of objects $d(i, j)$. Using hierarchical clustering, the clusters whose distance $D(A, B)$ is a minimum, are merged as one larger cluster. Similar combinations continue until the hierarchical tree is finally constructed.

Ward's Link: The cluster analysis technique, Ward's link, was proposed by [186]. The method works in a manner that the pair of clusters whose combination results in minimum increase in "information loss" are merged. This happens on every possible pair of clusters at each stage of the hierarchical clustering. "Information

loss" is defined by Ward in terms of an error sum of squares criterion, ESS.

$$ESS = \frac{N_A N_B}{N_A + N_B} \|\bar{X}_A - \bar{X}_B\|^2 \tag{4.4}$$

EES is the error sum of squares, $\bar{X}_A$, $\bar{X}_B$ are the centroids of clusters A and B, $\|\bar{X}_A - \bar{X}_B\|$ is the Euclidean distance between the clusters A and B, N_A and N_B are the number of objects in clusters A and B.

4.1.4 Clustering of SFRB data

Based on different distance measures and linkage criteria, four different cluster algorithms are applied on the 372 SFRB data points (each point was associated with 43 variable values), aiming to find the intrinsic groups underlying the SFRB data. After clustering, the type of each SFRB can be determined. Moreover, the clustering results can be displayed on a dendrogram, which allows visualization of the cluster structure of the data.

In addition, to evaluate the performance of the clustering method, regarding each generated cluster, precision and recall are used.

$$Precision = \frac{n}{m} \tag{4.5}$$

where n denotes, the number of observations that are correctly predicted to be in a particular cluster; m means the total number of observations in the same cluster.

$$Recall = \frac{n}{l} \tag{4.6}$$

where n represents the number of instances which are correctly predicted as belonging to a specific SFRB type in a cluster, l is defined as the number of SFRB of the same type by experts in the evaluated data.

4.2 Principal Component Analysis (PCA)

4.2.1 Rational

In many real-world classification problems, the relevant features (i.e. characterization variables) are often unknown prior to starting the investigation. A large number of features are therefore introduced in an attempt to better represent the underlying structure of a data set. Thus, many features, which are irrelevant and/or redundant with respect to a particular site, are recorded anyway, and subsequently analyzed. Redundant and irrelevant features often have adverse effects on data mining such as over-fitting and reduced classification accuracy, and are therefore undesirable. Effective feature transformation and feature selection techniques are thus applied to reduce the dimensionality. The objective of feature transformation is to project the data set in a new feature space with fewer dimensions.

The SFRB concept provides a holistic assessment of a water body or impoundment including its diffuse pollution control and ecological status, functional role, and flood control potential. In contrast to conventional assessment approaches, the SFRB assessment integrates hard engineering control variables including dam height and length with more holistic variables such as how highly engineered an SFRB appears, and whether the structure represents a significant barrier to aquatic or land animal passage (see Chapter 3). The large range of information gathered aims to greatly assist in addressing multi-disciplinary issues regarding

the management and use of a water body. However, some of the variables have direct or indirect relationships with each other, which means that dependency (or correlation) is always hidden among these variables. A key scientific question arises. How can one reveal the correlations among the variables, identify the important variables and subsequently remove the redundant ones to make SFRB characterization more effective and efficient? To weight and prioritize the variables, Scholz [159, 160] used the sum of all absolute correlations coefficients for one variable with all other variables multiplied by the mean confidence of each variable, which was finally named as priority point. It was assumed that the larger the priority point, the more important the variable was. However, this priority point failed to reflect the importance of variables since the method was based on assumptions that are not always true. Alternatively, Principal Component Analysis (PCA), projecting data into a new feature space, aims to reduce dimensionality by selecting the smallest number components that account for most of the variation in the original multivariate data [97]. Therefore, PCA provides a suitable way to reduce dimensions of SFRB data and thus to simplify SFRB system.

4.2.2 Related Work

To analyze the high-dimensional data sets, various dimension reduction methods have been published in the statistics, signal processing and machine learning literature, including independent component analysis [72], factor analysis [113], projection pursuit [84] and principal component analysis [75, 83]. Different techniques have specific advantages and drawbacks. More detailed information can refer to the technical report [45]. Among these methods, principal component analysis (PCA), invented by Karl Pearson [129], is the best (in the mean-square error sense) linear technique for dimensionality reduction [45]. Diverse

applications of PCA include data compression, image processing, visualization, exploratory data analysis, pattern recognition, predictions of time series, etc [178]. A detailed discussion of PCA can be found in textbooks [141, 67, 83].

Take practical applications for instance, PCA has been widely used in biomedical problems [69, 138]. Staniford-Chen and Heberlein [172] adopted this method to infer the best choice of thumb-printing parameters from data. PCA was also used to reduce dimensionality of trace data for intrusion detection [176, 168]. In addition, PCA was applied to detection and visualization of computer network attacks by Labib and Vemuri [97]. Furthermore, PCA was introduced into SFRB research to reveal the underlying data structure of retentions basins in Baden, Germany [161, 163]. Based on the loading plot of PCA on the first and second principal components, it could be seen that the variables that stayed isolated were independent or important variables. However, it is unclear whether there were more principal components that should be have been removed. Further investigation is necessary to determine whether all the variables removed were in fact redundant. Moreover, it is difficult to identify the important original variables since the components displayed on the plot are linear combinations of the original variables in a new space.

4.2.3 Principal Component Analysis Algorithm

PCA transforms the original correlated multivariate attributes to a set of mutually uncorrelated and orthogonal components (known as Principal Components), each of which is a particular linear combination of the original attributes. Principal components are estimated from the eigenvectors of the covariance matrix of the original variables. The principal components which account for most of the variance of the original data are used to summarize the original data. Therefore, it reduces data dimensionality with little loss of information.

Given a data matrix X consisting of N observations with M variables, supposing that each observation is expected be described with only L variables, $L < M$. The PCA algorithm is undertaken in five steps. First, the data are standardized to zero mean and unit standard deviation since the multiple variables are not on a comparable scale. Secondly, the covariance matrix (C) is calculated based on the standardized data.

$$u = \frac{1}{N}\sum_{n=1}^{N} X_n \tag{4.7}$$

where N is the number of observations, X is the original data matrix, u is the empirical mean of X.

$$C = \frac{1}{N}\sum (x - u_h)(x - u_h)^T \tag{4.8}$$

where u_h is a unit vector including 1 column and N rows, $x - u_h$ is the deviation from the empirical mean, $(x - u_h)^T$ is the conjugate transpose operator, C is the covariance matrix of standardized x.

The third step is to find the eigenvectors and eigenvalues of the covariance matrix. The eigenvalues and eigenvectors are paired.

$$V^{-1}CV = D \tag{4.9}$$

where C is the covariance matrix of the standardized data, D is the diagonal matrix of eigenvalues of C, Matrix V (with dimension of $M \times M$) represents the eigenvectors of the covariance matrix C.

In the fourth step, the columns of the eigenvector matrix V and eigenvalue matrix D are sorted in order of decreasing eigenvalue. The eigenvalues represent the distribution of the source data's energy among all eigenvectors. The final step is to determine the smallest L subset of the eigenvectors which account for a certain high proportion (like 90%) of the cumulative energy of all the eigenvectors.

$$\frac{\sum_{q=1}^{L} D[q,q]}{\sum_{q=1}^{M} D[q,q]} \geq 90\% \tag{4.10}$$

where $1 \leq L \leq M$, $\sum_{q=1}^{L} D[q,q]$ means the cumulative eigenvalues of the first L eigenvectors, $\sum_{q=1}^{M} D[q,q]$ represents the cumulative eigenvalues of all the M eigenvectors.

The above equation indicates that the first L principal components account for the highest variance of the multivariate data and thus PCA reduces the original dimensions from M to L.

However, the selected principal components are the important components in the new space rather than in the original space. Therefore, it is hard to interpret which original variables are relatively important. Equation 4.11 is used in order to evaluate the contribution of the original variables to the selected principal components.

$$Contribution(q) = \sum_{p=1}^{L} D_p V_p^q \tag{4.11}$$

where q is q^{th} variable, L is the selected number of principal components, $q = 1, 2, \cdots, L$, D_p means the p^{th} eigenvalue, V_p^q is the q^{th} value of the p^{th} eigenvector of the data.

By integrating the eigenvalues and the eigenvectors, the contributions were calculated and sorted into descending order. The larger the contribution, the more important the variable is for all the L principal components.

4.2.4 PCA on SFRB Data

PCA is applied to the Scottish SFRB data set. It aims to reduce dimensions of SFRB system by projecting the original data into a new feature space with fewer dimensions. Another objective is to investigate each variable's contribution to the principal components and thus to identify the key variables and reduce the redundant ones. The analysis was conduced by using the Matlab software package [133].

4.3 Self-organizing Map (SOM)

4.3.1 Rational

As stated in the last section (cf. Section 4.2), the variables in high-dimensional data are often correlated, which can be analyzed by using PCA. The principal components in new space can reduce the dimensionality of the data set effectively. However, the components generated in the new space, consisting of linear combinations of the original variables, are very difficult to interpret. So, PCA is not the good way to reveal the correlations among the original variables of SFRB. To address this problem, Self-organizing Map (SOM) is introduced. In SOM, the original data set is represented as a smaller set of prototype vectors (or called feature vectors) [185]. The prototypes with close relationships will tend to group together and form clusters. In contrast with PCA, SOM has many

advantages [106, 108]. One major benefit of SOM is to allow visual presentation and interpretation of the input patterns in a two-dimensional feature map. It brings deep insight into SFRB data patterns by vector quantization. In addition, through competitive learning, the similar SFRB group together in the feature map and the missing values of variables can thus be effectively estimated by searching for the best matching units (BMUs).

4.3.2 Related Work

Artificial neural network (ANN) was initially developed in 1940s by McCulloch and Pitts [115]. It is a mathematical model that is inspired by the structure and/or functional aspects of biological neural networks. ANN has been widely used in such diverse areas as engineering, medicine, computer science, psychology, neuroscience, physics, mathematics and biology [123]. In contrast to other artificial neural networks, the SOM is an unsupervised learning algorithm, that captures the topological properties of the input data according to the intrinsic data structure, and little therefore needs to be known about the structure of the input data [2]. Moreover, SOM allows for intuitive visualization of correlations between features by vector quantization and for mapping multidimensional data into two-dimensional space. SOM has been widely studied and applied in various pattern recognition tasks, practical speech recognition, process control, etc [93]. Due to the robustness of the method, there has been a steady increase in the number of applications in Meteorology and Oceanography [107], music recommendation systems [184] and in water resources [6, 158, 99, 85].For example, the SOM model was successfully used to elaborate heavy metal removal mechanisms and to predict heavy metal concentrations in experimental constructed wetlands treating urban runoff [99].

4.3.3 SOM Algorithm

SOM is a neural network model that is trained using competitive learning [93, 94]. Initially, the input data need to be normalized, so that each component has unit variance. SOM involves two processes: training and mapping. Training develops a feature map to respond to the input patterns of the data by competitive learning, also called vector quantization. During mapping, a new input vector is projected onto the feature map by finding the node with the closest weight vector to the vector of interest, and assigning the map coordinates of this node to the vector.

First, to construct the neural network to represent the input patterns, the number of neurons (also called nodes) is selected by the heuristic Equation (4.12).

$$M \approx 5\sqrt{n} \tag{4.12}$$

where M is the number of neurons; and n is the number of data samples.

Each neuron i is represented by an n-dimensional weight vector $m_i = [m_{i1}, \cdots, m_{in}]$. The weight vectors are initialized with random values. When an input vector is fed into the network, its Euclidian distances to all weight vectors are computed. The neuron with the weight vector most similar to the input vector is chosen as the best matching unit (BMU). The BMU and its topological neighbours are then moved closer to the input vector by updating their weight vectors. The update rule is governed by Equation (4.13).

$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci}(t)[x(t) - m_t(t)] \tag{4.13}$$

where $m(t)$ is the weight vector indicating the output unit's location in the data

space at time t; $\alpha(t)$ is learning rate at time t; h_{ci} is the neighbourhood function centred in the winner unit c at time t; and $x(t)$ is input vector drawn from the input data set at time t.

Through the competitive learning process, the SOM forms a feature map where similar samples are mapped closely together and dissimilar samples further apart. The various input patterns are thus automatically represented by the feature map. The quality of the learning is measured by the quantization error (QE) and the topographic error (TE). The QE is the mean distance between each data point and its BMU, and TE represents the proportion of all data for which the first and second BMU are not adjacent with respect to the measurement of topology preservation [94].

After training the SOM with the input samples through competitive learning, the created feature maps represent the various patterns of training data, and thus the relationships among variables are visualized. In addition, SOM can also be used to predict the missing values of a new input vector by searching for its BMU. Considering the correlations among variables, the missing values of one variable can be predicted based on its highly related variables.

4.3.4 Predictions Based on SOM

The application of the SOM for prediction purposes is illustrated in Figure 4.1. Initially, the whole SFRB data set (372 samples) is split into two subsets, each of which consists some samples of each SFRB type. This is to ensure that the pattern of each SFRB type can be trained and tested, otherwise the SOM model will become incomplete. One subset of 188 samples is used for SOM training, and the other subset consisting of the remaining 184 samples is used to test the predictive and classification abilities of the SOM model. Testing of the SOM's

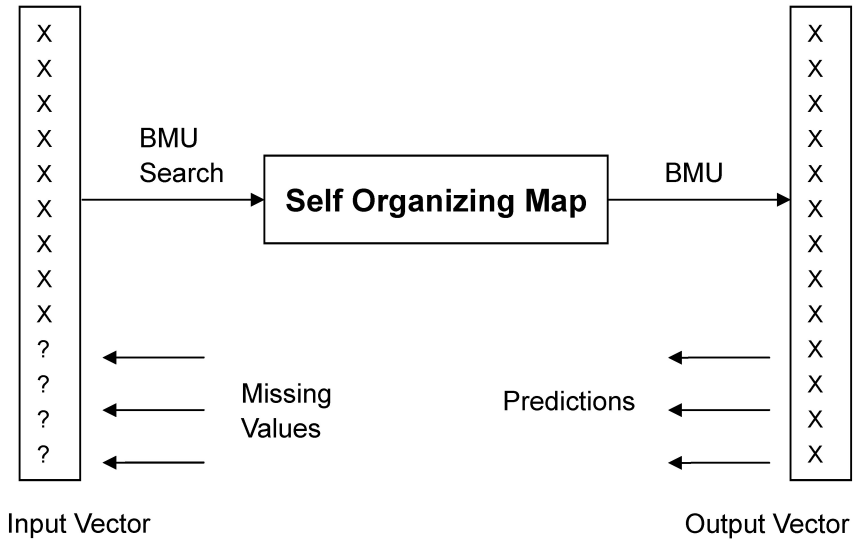


Figure 4.1: Prediction of the missing components of the input vectors using self-organizing map modeling.

neural network was undertaken by removing the variable of interest from the data set. The missing values result in a depleted vector within the overall SOM, and the model resolves this by identifying the BMU for the depleted vector. The values for the missing variables are then obtained by their corresponding values in the BMU [149]. Similarly, it can be used to predict the missing types of SFRB. Finally, the QE and TE are calculated to evaluate the quality of the SOM predictions. All statistical analysis are performed using the standard software package Matlab 7.0 (http://www.mathworks.co.uk/).

4.4 Feature Selection

4.4.1 Rational

In many fields of research and practice solutions to problems of a highly multidimensional nature are sought. The traditional way to deal with such

multidimensional data, the traditional way is to use data transformation(e.g. PCA) to map the data set into a lower-dimensional space with little information lost. However, as mentioned in Section 4.2, the principal components (the extracted novel features) obtained by PCA are linear combinations involving all original features, which are difficult to interpret. Therefore, PCA is not an effective way to identify the important variables of SFRB.

To deal with such problems, feature selection is proposed. In contrast to feature transformation, feature selection does not alter the original representation of the variables, but merely selects a subset of these. Feature selection techniques preserve the original semantics of the variables, and someone with expert knowledge of these variables can therefore interpret these accurately. The general concept is to identify the most relevant attributes for the concept at hand, thus facilitating assessment of the data. Feature selection therefore selects an optimal subset of the available variables, achieving the best classification performance with the fewest variables. Redundant and unimportant variables are thus excluded from the classification, leading to an improved overall result. The reduction of feature space speeds up learning algorithms and leads to benefits in terms of classification accuracy and interpretability of the classification result. Identification of irrelevant variables ensures that these can be eliminated from the classification system, thus reducing the time and costs associated with data collection and analysis. In particular, feature selection helps to gain a deeper insight into the underlying data structure. It is therefore a powerful technique for reducing highly multi-dimensional data sets to manageable levels. This approach makes SFRB assessment more rapid, efficient and cost effective, providing the EU member states with a rapid tool to implement the EU Flood Directive, and supporting engineers and planner regarding design, maintenance and management.

4.4.2 Related Work

During the last several decades, various feature selection methods have been studied comprehensively [104, 30]. Feature selection involves searching through various feature subsets and evaluating each of these subsets using appropriate criteria. The most popular search strategies are 'greedy' sequential searches through the feature space, either forward or backward. The evaluation criteria are roughly classified into filter and wrapper methods [59]. A filter model [103] examines the intrinsic properties of the data, such as the simple statistics computed from empirical distributions, to select and evaluate feature subsets without involving any classification algorithm. A wrapper method [92] requires a specified classifier and uses the wrapper's performance as the evaluation criterion. It involves finding the optimal subsets of features, which achieve the highest classification accuracy. Therefore, the filter model has the independent criteria and higher computational efficiency than the wrapper method. Consequently, many filter models have been proposed, including Information Gain [62], Relief [91], Minimal-redundancy-maximal-relevance (MRMR; [130]), Focus [3] and Correlation-based Feature Selection [63]. These feature selection techniques have been widely applied in many domains such as text categorization [48], image retrieval [147], magnetic resonance images classification, bioinformatics [151], land cover classification for a semi-arid environment [13] and human gait recognition [58].

4.4.3 Feature Selection Algorithms

Among different evaluation criteria for filter feature selection, information-theoretic methods seem to be more comprehensively studied. The main reason is that information entropy offers a good measurement to quantify the uncertainty of a feature. The information entropy is intuitive, and generally results in good

classifier performance, which is independent of the type of the classifier. Two established information-theoretic filter techniques (Information Gain and Mutual Information) are applied. Relief is also included in the analysis, because it is an effective evaluation criterion based on the local data distribution. These algorithms are described in detail below.

Information Gain: Information Gain is employed as a relevance-criterion to measure the number of bits of information obtained for the class prediction by knowing the presence or absence of an attribute [62]. This technique is highly efficient to compute, and it is also not restricted to linear correlations, but captures arbitrary dependencies between features. For a given attribute A with respect to the class attribute C, the Information Gain is the reduction in uncertainty about the value of C when the value of A is known. The uncertainty about the class attribute of C is measured by its entropy $H(C)$. Therefore, the entropy of the class C before and after observing the attribute A is given by Equations (4.14) and (4.15).

$$H(C) = -\sum_{c \in C} p(c) \log_2 p(c) \tag{4.14}$$

where $H(C)$ is the entropy of the class attribute C, and $p(c)$ is the probability mass function of the outcome c.

$$H(C|A) = -\sum_{a \in A} p(a) \sum_{a \in C} p(c|a) \log_2 p(c|a) \tag{4.15}$$

where $H(C|A)$ is the conditional entropy of C for a given A, $p(c)$ is the probability mass function of the outcome c, and $p(c|a)$ is the conditional probability of c for a given a.

Since the reduction of the entropy of the class after knowing the attribute A reflects the additional information about the class provided by the attribute, Information Gain captures the interestingness of variables for the class concept [135]. Formally, the Information Gain $I(C;A)$ is defined as shown in Equation (4.16).

$$I(C;A) = H(C) - H(C|A) \tag{4.16}$$

where $I(C;A)$ is the information gain, $H(C)$ is the entropy of the class attribute C, and $H(C|A)$ is the conditional entropy of C given A.

Mutual Information: As a measure of relevance and redundancy among features, Mutual Information of two random variables is a quantity that measures the mutual dependence of the two variables [40]. A heuristic minimal-redundancy-maximal-relevance (mRMR) framework can be used to select promising features for both continuous and discrete data sets [130]. The maximal-relevance criterion (Max-Relevance with class target) searches the attribute subset S, which approximates $D(S,C)$ in Equation (4.17).

$$D = max\{\frac{1}{|S|}\sum_{A_i \in S} I(A_i;C)\} \tag{4.17}$$

where D is maximal mutual information between each attribute A_i and class C, S is the selected attribute subset, $|S|$ is the number of attributes in S, and $I(A_i;C)$ is the mutual information between attribute A_i and class C.

It is likely that features selected according to Max-Relevance could result in redundancy; i.e. the dependency among these features could be large. When two features are highly dependant on each other, the respective class-discriminative

power does not change much, if one of them were removed. Therefore, the minimal redundancy (Min-Redundancy) condition is added to select mutually exclusive features (Equation (4.18)).

$$R = min\{\frac{1}{|S|^2}\sum_{A_i,A_j \in S} I(A_i, A_j)\} \tag{4.18}$$

where R is minimal mutual information among selected attributes, S is the selected attribute subset, $|S|$ is the number of attributes in S, and $I(A_i, A_j)$ represents the mutual information between the two attributes A_i and A_j.

The criterion combining the above two constraints is called minimal-redundancy-maximal-relevance (mRMR). Therefore, to optimize D and R simultaneously, the objective is to maximize the value of (D, R) as shown in Equation (4.19).

$$\Phi = max(D - R) \tag{4.19}$$

where Φ is defined as the operator, combining D and R, which represent the dependency and redundancy of a feature subset on the target class, respectively.

In practice, an incremental search method is used to find the near-optimal features defined by $\Phi(\cdot)$. This is done by optimizing the condition expressed in Equation (4.20).

$$max_{A_j \in A - S_{m-1}}[I(A_j; C) - \frac{1}{m-l}\sum_{A_i \in S_{m-l}} I(A_j; A_i)] \tag{4.20}$$

where A is the whole attributes, S_{m-l} is a attribute subset with $m - l$ attributes that have been obtained, $I(A_j; C)$ is the mutual information between the attribute

A_j and the target class C, and $I(A_j; A_i)$ is the mutual information between the two attributes A_i and A_j.

Relief: Relief is a classical instance-based attribute selection scheme introduced by Kira and Rendell [91], and enhanced by Kononenko [95]. The key idea of Relief is to estimate attributes according to how well their values distinguish among instances of different classes that are near each other [95]. For that purpose, Relief for a given instance searches for its two nearest neighbours; one is taken from the same class (called nearest hit H) and the other from a different class (called nearest miss M). For each attribute, it calculates the relevance scores and updates its value according to Equation (4.21).

$$W(A) = W(A) - diff(A, X, H)/m + diff(A, X, M)/m \tag{4.21}$$

where $W(A)$ represents the relevance scores for any attribute A, $diff(A, X, H)$ is the difference between the values of attribute A for the two instances X and H, and m is the number of instances sampled.

This process is repeated for a user-specified number of instances m. The rationale is that a useful attribute should differentiate between instances from different classes and have the same value for instances from the same class [62].

4.4.4 Classification Algorithms

Briefly, classification is defined as learning a function, model or other method from a subset of the data objects to assign a data object to one of several predefined classes. Firstly at the training stage, some amount of instances with known class labels are used to learn the data information. Then in the test phase,

the classifier predicts the class label of unlabeled instances based on the learned information. To assess the effectiveness of the feature selection methods, four benchmark classification algorithms with very different algorithmic paradigms are used: Support Vector Machine (SVM), K-Nearest Neighbour (KNN), C4.5 Decision Tree (J48) and Naïve Bayes (NB). They are described respectively as follows.

Support Vector Machine (SVM): SVM is a widely used and promising tool for data classification. Its basic idea is to construct a separating hyperplane between the training instances of both classes. Among all possible hyperplanes, the one with the maximum margin between classes is selected [22]. Given training vectors $x_k \in R^n (k = 1, \cdots, m)$ in two classes, and a vector of labels $y \in R^m$ such that $y_k \in \{1, -1\}$, SVM solves a quadratic optimization problem (Equation (4.22)). For any testing instance x, the decision function (predictor) has the form outlined in Equation (4.23).

$$min_{\omega,b,\varepsilon} \frac{1}{2}\omega^T\omega + C\sum_{k=1}^{m}\varepsilon_k \tag{4.22}$$

subject to

$$y_k(\omega^T\phi(x_k) + b) \geq 1 - \varepsilon_k, \varepsilon_k \geq 0, k = 1, \cdots, m$$

where ω is a normal vector, b is a scalar and ε_k are non-negative variables, C is a penalty parameter on the training error, y_k is the class label, and $\phi(\cdot)$ is a map function to transfer the training data into a higher dimensional space.

$$f(x) = sgn(\omega^T\phi(x) + b) \tag{4.23}$$

where $f(x)$ is the prediction function, $sgn(\cdot)$ is a symbol function, ω^T is the permutation of normal vector ω, $\phi(\cdot)$ is a map function, and b is a scalar.

It is practical for the Kernel function (Equation (4.24)) to be used to train the SVM.

$$k(x, y) = \phi(x) \cdot \phi(y) \tag{4.24}$$

where the linear kernel function $k(x, y)$ used in this study and $\phi(\cdot)$ is a map function.

K-Nearest Neighbour (KNN): A simple k-nearest neighbour classification algorithm is used by setting k equal to three. The distance metric that has been used is the Pearson correlation coefficient. The k-nearest neighbour is a supervised learning algorithm based on instances [1]. It simply stores the training data and postpones the generation until an instance must be classified. Given an instance, its k closest neighbours are found in terms of the Pearson correlation coefficient, and then its label value is determined by the k neighbours using the majority vote manner principle. In this study $k = 3$.

C4.5 Decision Tree: C4.5 is a statistical classification algorithm used to generate a top-down decision tree developed by [135]. Each node of the tree is constructed by finding the attribute of the data that most effectively splits its set of samples into subsets. The attribute with the highest normalized information gain is chosen to make the decision to split the data. The C4.5 algorithm then generates smaller sub-lists in a recursive way. Despite only one feature being chosen at a time, this depends on previous results. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA (http://www.cs.waikato.ac.nz/ml/weka) data mining tool.

Naïve Bayes: The applied Bayesian classifier is a simple probabilistic classifier based on Bayes' theorem (from Bayesian statistics) with strong independence assumptions [143]. Only the variances of the variables for each class need to be determined, and not the entire covariance matrix. Thus, an advantage of the Naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Due to the precise nature of the probability model, Naïve Bayes classifiers can be trained very efficiently in a supervised learning setting.

4.4.5 Identifying Key Variables of SFRB

In this study, in order to find the most important variables of SFRB, the feature selection approach is used. To avoid the potential influence of special algorithmic characteristics of a single method on the result, different feature selection algorithms and different classifiers are combined. In general, feature selection approach comprises four phases (see Figure 4.2). First, the entire data set is split randomly into 10 folds. At each time, any nine folds are selected for training the feature selection and classification algorithms, while the remaining fold is used for testing. In the second phase, three popular feature selection algorithms (Information Gain, Mutual Information and Relief) are applied on the SFRB data set, obtaining three different feature sequences according to the priority of the degree of relevance for the SFRB types. In the third phase, one uniform sequence of variables is determined. Since the feature rankings generated by the three feature selection algorithms are different, the sequence number of each variable in each feature sequence is counted and these are subsequently summed up. Then, the sum of their sequences is ranked to obtain the final and uniform order of the 40 variables. Finally, different numbers of the ranked features are passed to a classifier to further assess the effectiveness of the selected

features by using the classification accuracy as the decision criteria. Four popular classifiers are used in this study: Support Vector Machine (SVM), k-Nearest Neighbours (KNN), C4.5 Decision Tree (J48) and Naïve Bayes (NB). A 10-fold-cross-validation [174] is performed during the process of classification.

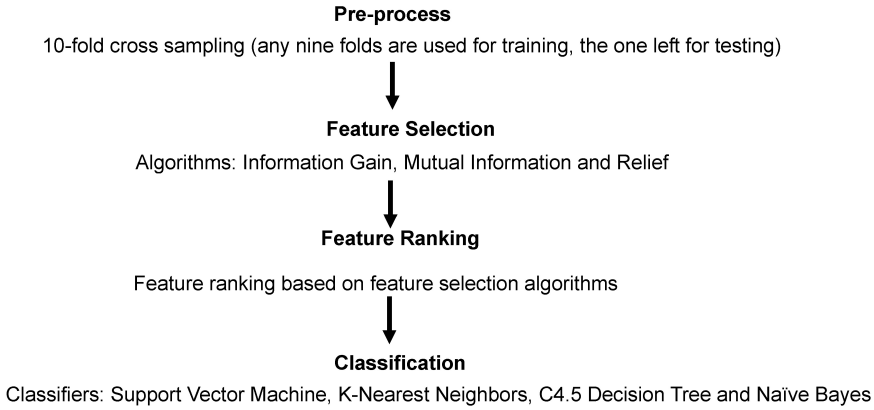


Figure 4.2: Framework of the assessment approach.

4.5 Multi-label Classification

4.5.1 Rational

Currently, classification models of SFRB based on correlations between characteristic variables [159] and traditional supervised learning (cf. Section 4.4) have been developed to distinguish SFRB types. Classification is a primary technique used to predict class membership for data instances involving two steps: training and testing. In the training step, the classifier learns the different class concepts on training data with known class labels. In the test step, the classifier predicts the class label of unlabeled instances based on the learned information. However, all these traditional methods are limited to assigning only one label (SFRB type) to one SFRB site, which means that one SFRB is predicted as one specific (i.e.

the only one) SFRB type. In fact, one basin usually has multiple functions and fall into more than one SFRB category.

For instance, Figure 4.3 shows a picture of Harlaw Reservoir ($55.87°N$, $3.32°W$), which not only supplies water for Edinburgh (SFRB type 2) but also plays an important role in flood control (SFRB type 1) after the change of the management strategy in 2010. During the periods of high rainfall, sufficient space in the reservoir is left to collect unexpected flood water. The reservoir holds back water running off the Pentland Hills (South-west of Edinburgh) and releases the runoff gradually, avoiding floods downstream. Therefore, some existing drinking water reservoirs also perform flood control functions, which will reduce costs for new flood defence constructions.

Another challenge of traditional SFRB classification is data training. In reality, each SFRB is likely to perform different functions. However, traditional classifiers cannot learn multiple functions of SFRB, but only the predominant function of a SFRB. This leads to the incomplete learning of the pattern of the data set. Consequently, it is impossible to predict SFRB types effectively by using traditional classification. For instance, the SFRB classification in [163], which assumed one SFRB only belonged to one type, achieved accuracy as low as 60%. The inaccurate results of the traditional classification models will cause conflicts and misunderstandings of SFRB among various stakeholders such as designers, planners and managers. To better understand and capture the functions of SFRB and improve the development and management of SFRB in the context of flood risk management planning, the multi-label classification model is thus introduced in this study.

Figure 4.3: Harlaw Reservoir($55.87°N$, $3.32°W$) located in the Pentland Hills near Edinburgh.

4.5.2 Related work

Traditional classification is concerned with learning from a set of examples that are associated with a single label l from a set of labels L, $|L| > 1$. If examples are associated with a set of labels, this is called multi-label classification, which refers to the classification problems where an instance can belong to more than one category simultaneously [180].

Various multi-label classification approaches have been extensively studied. Initially, this approach originated from an investigation of text categorization problems, where each document may belong to several labels (topics) simultaneously [114]. Joachims [81] constructed a set of binary support vector machine SVM classifiers, training each possible class versus the remaining ones and assigning a real value to each class to indicate the class relationship. However, this method

did not address multi-label training models and specific testing criteria. McCallum [114] proposed a generative approach in which a model was trained by Expectation Maximization (EM); i.e. selection of the most probable set of labels from the set of possible classes. However, this generative model was limited to specific text applications. Schapire and Singer [157] proposed BoosTexter, extending AdaBoost to handle multi-label text categorization. The algorithm was somewhat inefficient due to a high space complexity and time-per-round complexity. Recently, multi-label learning has been widely applied in many complex real world problems ranging from bioinformatics [25, 37, 193, 197], music [102, 179], directed marketing [195], video annotation [134, 170] to semantic scene classification [14]. For bioinformatics, the C4.5 algorithm was adapted by Clare and King [25], who modified the definition of entropy to include multi-label data of gene expression. Choosing a decision tree as the baseline algorithm, they only aimed to learn the symbolic rules. However, it was not a complete classification. Elisseeff and Weston [37] proposed a multi-label kernel method for Yeast gene functional classification, minimizing the ranking loss (see also Section 4.5.4). Zhang and Zhou [193] extended the back-propagation neural network algorithm for multi-label learning (BP-MLL). They introduced a new error function that takes multi-label learning into account. In addition, Boutell et al. [14] proposed a Multi-label Support Vector Machine (MLSVM) framework by decomposing the multi-label learning problems into multiple independent binary classification problems. They presented an effective cross-training strategy, the C-Criterion in testing and two new evaluation metrics. Zhang and Zhou [194] proposed a lazy learning approach to multi-label learning named Multi-label k-Nearest Neighbours (MLKNN), extending the KNN algorithm with a Bayesian approach. It uses the maximum a posteriori (MAP) principle to determine the label set of the test instance, based on prior and posterior probabilities for the frequency of each label within the k nearest neighbours.

4.5.3 Multi-Label Classification Algorithms

In traditional supervised learning, an instance is associated with a class label. Formally, let X be the domain of instances, Y is the set of labels, the goal is to learn a optimal classifier $h : XY$ from a given data set $\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$, where $x_i \in X$ is an instance and $y_i \in Y$ is the known label of x_i.

In contrast, for multi-label problems, each instance is associated with multiple class labels. Formally, let X denote the set of instances and Y the set of class labels. Then the task is to learn a classifier hml: X2Y from a given data set $\{(x_1, Y_1), (x_2, Y_2), \cdots, (x_m, Y_m)\}$, where $x_i \in X$; $Y_i \in Y$ is a set of labels $\{y_i^1, y_i^2, \cdots, y_i^k\}$, $y_i^k \in Y$ $(k = 1, 2, \cdots, l_i)$ and l_i is the number of labels in Y_i.

Currently, various multi-label classification algorithms have been proposed. Here, three popular algorithms MLSVM, MLKNN and BP-MLL are applied for multi-label classification of the SFRB data sets from Scotland and Baden. The reason for selecting these algorithms was that these methods are widely used and are proven to have good performance. Additionally, traditional supervised classifiers SVM, KNN and BP are selected for further comparison. During the training phase, one-leave-out cross-validation is performed. To avoid the sensitivity of parameter setting affecting classification results, several parameters are selected in a heuristic way as suggested by these algorithms.

Multi-Label Support Vector Machine (MLSVM): Based on SVM (Section 4.4.4), Boutell et al. [14] applied multi-label learning techniques (MLSVM) to scene classification by converting the multi-label learning problem into multiple independent binary classification problems. Cross-training (also called MODEL-x) is used where multi-label data are trained more than once. If a basin belongs, for example, to SFRB type 1 and 3 simultaneously, for learning the concept of type 1, the SFRB is selected as the positive instance. It is also further treated as the

positive instance when learning the concept of type 3. For handling multi-class problems, the one-versus-all strategy is applied, where each class is compared with all other classes. Although the experiments indicate the effectiveness of this strategy, its performance may also be affected by other factors, which are not considered here, such as small sample size and unbalance data training. For more detail, please refer to [142] for experimental studies on different multi-class classification algorithms, although the problem of which strategy can yield better practical performance is still unresolved.

After the classifier learns each concept of the classes during the testing phase, given a test instance, it outputs a score for each class. The magnitude of the score indicates the degree of membership in the corresponding class. To generalize the one-versus-all approach with respect to multi-label classification, a labeling criterion (i.e. the C-Criterion) is used to determine the labels of an instance in the testing phase. For this criterion, among all the scores for an instance and if the top M are close enough, then the corresponding classes are considered as the labels for that instance. To judge if the scores are close enough or not, the maximum a posteriori principle is introduced to determine the threshold for selecting the M [14]. Here, to reduce the impact of the parameter C on the training error, five setting $C = \{0.2, 0.4, 0.6, 0.8, 1.0\}$ are used during the experiment.The results are averaged over all different parameters.

Multi-Label K-Nearest Neighbour (MLKNN): Derived from the popular K-Nearest Neighbour (KNN) algorithm, a lazy learning algorithm named Multi-Label K-Nearest Neighbour (MLKNN) has been proposed [194]. Firstly, for each test instance, MLKNN defines its k-nearest Neighbours in the training set. In this study, $k = \{6, 8, 10, 12, 14\}$ were used to mitigate the impact of classification results in terms of the parameter of number of neighbours. Then, based on prior and posterior probabilities for the frequency of each label within the k-nearest

Neighbours, the maximum a posteriori principle is utilized to determine the label set for the test instance.

Given instance $x_i(i = 1, 2, \cdots, m)$ and its associated label set $Y_i \in Y$, based on label sets of its k neighbours, first, the membership counting vector can be defined as in Equation (4.25):

$$\boldsymbol{C}_{x_i}(l) = \sum_{a \in N(x_i)} \boldsymbol{y}_a(l), l \in Y \tag{4.25}$$

where $\boldsymbol{C}_{x_i}(l)$ counts the number of neighbours of x_i belonging to the $l-th$ class; $N(x_i)$ denotes the set of k-nearest neighbours of x_i identified in the training set; $\boldsymbol{y}_a$ means category vector for instance a, where the $l-th$ component $\boldsymbol{y}_a(l)$ equals 1 if $l \in Y_i$ and 0 otherwise.

Given an instance $x_i(i = 1, 2, \cdots, m)$, the prior probabilities $P(H_1^l)$ $(l \in Y, b \in \{0, 1\})$ and posterior probabilities $P(E_j^l|H_1^l)$ $(j \in \{0, 1, , k\})$ are estimated by Equation (4.26) to Equation (4.28).

$$P(H_1^l) = (s + \sum_{i=1}^{m} \boldsymbol{y}_{x_i}(l))/(s \times 2 + m) \tag{4.26}$$

$$P(H_0^l) = 1 - P(H_1^l) \tag{4.27}$$

$$P(E_j^l|H_1^l) = (s + c[j])/(s \times (k+1) + \sum_{p=0}^{k} c[p]) \tag{4.28}$$

where H_1^l denotes the event that x_i has the label l, while H_0^l is the event that x_i has not label l; E_j^l $(j \in \{0, 1, \cdots, k\})$ represents the event that j instances among

the k-nearest neighbours have label l; $P(H_b^l)$ and $P(E_j^l|H_b^l)$ are prior and posterior probabilities of the events; m is the number of instances; the input argument s is a smoothing parameter used to control the strength of uniform prior (1 in default); $c[j]$ counts the number of training instances with label l whose k nearest neighbours contain exactly j instances with label l.

For each test instance t, k-nearest neighbours $N(t)$ were identified in the training set. On the basis of the membership counting vector $\boldsymbol{C}_t$, the category vector $\boldsymbol{y}_t$ is determined using the maximum a posteriori principle as shown in Equation (4.29).

$$\boldsymbol{y}_t(l) = argmax_{b\in\{0,1\}} P(H_b^l|E^l_{\boldsymbol{C_i}(l)}), l \in Y \tag{4.29}$$

Using the Bayesian rule, the MAP equation can be rewritten as expressed in Equation (4.30).

$$\boldsymbol{y}_t(l) = argmax_{b\in\{0,1\}} P(H_b^l)P(E^l_{\boldsymbol{C_i}(l)}|H_b^l) \tag{4.30}$$

where $\boldsymbol{y}_t(l)$ denotes the $l-th$ component of the category vector for t; $P(H_b^l)$ and $P(E_j^l|H_b^l)$ are the prior and posterior probabilities respectively, which can be calculated from Equation (4.26) to Equation (4.28).

$\boldsymbol{r_t}$ is a real-valued vector calculated to rank labels in Y and corresponds to the posterior probability $P(H_1^l|E^l_{\boldsymbol{C_t}(l)})$ which are described in Equation (4.31).

$$\boldsymbol{r_t}(l) = P(H_1^l|E^l_{\boldsymbol{C_t}(l)}) = (P(H_1^l)P(E^l_{\boldsymbol{C_t}(l)}|H_1^l))/\sum_{b\in\{0,1\}P(H_b^l)P(E^l_{\boldsymbol{C_t}(l)}|H_b^l)} \tag{4.31}$$

where H_1^l denotes the event that t has the label l; $E^l_{\boldsymbol{C_t}(l)}$ represents the event that, among the k-nearest neighbours, there are $\boldsymbol{C_t}(l)$ instances have label l; $P(H_1^l)$ and $P(E^l_{\boldsymbol{C_t}(l)}|H_b^l)$ are the prior and posterior probabilities of the events, respectively.

Back-Propagation for Multi-Label Learning (BP-MLL): The Back-Propagation (BP) model is a type of supervised feed-forward neural network. Neural networks are massively parallel interconnected networks of simple (usually adaptive) elements. Their hierarchical organizations interact with the objects of the real world in the same way as biological nervous systems do. Zhang and Zhou [193] extended the traditional multi-layer feed-forward neural network to the multi-label classification problem. Figure 4.4 shows the structure of BP learning which works in iterative steps [80]. They arranged neurons in multi-layers, of which the first layer takes inputs, the last layer produces the outputs and the middle one is called the hidden layer. The number of the hidden neurons = $\gamma\times$ input dimensionality, where the input dimensionality equals to the number of variables (43 in this study) and γ is a bias parameter of the feed-forward networks. To reduce the sensitivity of the parameter γ, the authors set it as 0.5, 1.0, 1.5, 2.0 and 2.5 in the experiments. To capture the characteristics of multi-label learning, a classical back-propagation algorithm [148], which uses gradient descent, was employed to minimize the global error function (see Equation (4.32)). The labels belonging to an instance should be ranked higher than those not belonging to that instance.

$$E = \sum_{i=1}^{m} E_i = \sum_{i=1}^{m} \frac{1}{|Y_i||\bar{Y}_i|} \sum_{(k,l)\in Y_i\times\bar{Y}_i} exp(-(c_k^i - c_l^i)) \tag{4.32}$$

where $\frac{1}{|Y_i||\bar{Y}_i|}\sum_{(k,l)\in Y_i\times\bar{Y}_i} exp(-(c_k^i - c_l^i))$ defines the error of the network on the $i-th$ multi-label training example (x_i, Y_i); $\bar{Y}_i$ is the complementary set of Y_i in Y and $|\cdot|$ measures the cardinality of a set; $c_k^i - c_l^i$ $(k \in Y_i, l \in \bar{Y}_i)$ means the

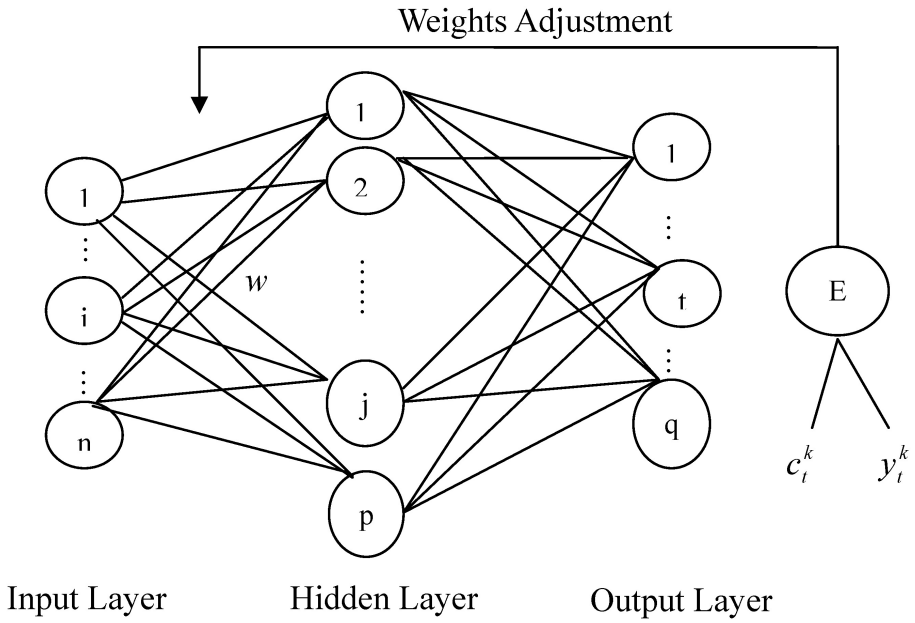


Figure 4.4: Typical BP network architecture.

difference between outputs of the network on one label belonging to x_i and one label not belonging to it.

The general error of the $j-th$ output unit can be defined as shown in Equation (4.33):

$$d_j = \begin{cases} (\frac{1}{|Y_i||\bar{Y_i}|}\sum_{l\in Y_i} exp(-(c_j - c_l)))(1+c_j)(1-c_j), & j \in Y_i \\ (-\frac{1}{|Y_i||\bar{Y_i}|}\sum_{k\in Y_i} exp(-(c_k - c_j)))(1+c_j)(1-c_j), & j \in \bar{Y_i} \end{cases} \tag{4.33}$$

4.5.4 Evaluation Measures

In the single-label system, a class prediction is either correct or incorrect and the standard evaluation metrics include precision, recall, accuracy and F-measure [165]. Here, for comparison, the accuracy is used, which is defined as follows:

$$Accuracy = \frac{n}{m} \tag{4.34}$$

where n denotes the number of instances that are correctly predicted; m means the number of instances in the evaluated data set.

In contrast, the evaluation of classification methods for multi-label data requires more complicated measures, because a result could be fully correct, partly correct or fully wrong. For the multi-label data, various evaluation measures have been studied [180, 181]. In this work, the following widely used evaluation metrics proposed by Schapire and Singer [157] are applied.

For the formal description of these metrics, let $D = \{(x_i, Y_i), i = 1, 2, \cdots, m\}$ denotes a multi-label evaluation data set and let $Y = \{\lambda_j : j = 1, \cdots, Q\}$ be the set of labels, $Y_i \subseteq Y$. Given an instance x_i, the predicted set of labels by a multi-label classifier is denoted as Z_i and the predicted rank of a label λ is denoted as $r_i(\lambda)$. Hamming Loss (Equation (4.35)) is based on the multi-label classifier, while all other metrics are defined based on multi-ranking methods. *Hamming Loss* evaluates how many times an instance-label pair is classified incorrectly. The smaller the value is, the better the performance of the classifier.

$$Hamming - Loss = \frac{1}{m} \sum_{i=1}^{m} \frac{|Y_i \Delta Z_i|}{|Q|} \tag{4.35}$$

where m is the number of the instances in the evaluated data set; Q is the size of the set of labels. Given an instance x_i, Y_i and Z_i stand for the actual set of labels and the predicted set of labels, respectively; Δ stands for the symmetric difference of two sets.

One-error (Equation (4.36)) evaluates how many times the top-ranked label does

not appear in the set of ground truth labels. The smaller the value of one-error is, the better the performance.

$$One - error = \frac{1}{m}\sum_{i=1}^{m}\delta(argmax_{\lambda \in Y} r_i(\lambda)) \tag{4.36}$$

where m is the number of the instances in the evaluated data set; $r_i(\lambda)$ represents the rank of a label λ predicted by the ranking method; Y stands for the set of labels; $\delta(\lambda)$ equals to one, if $\lambda \notin Y_i$, otherwise it equals to zero.

Coverage, formalized as below, is designed to assess how far one must go, on average, down the list of labels to cover all the possible labels that the instances actually belong to. The smaller the value of *Coverage* is, the better the performance.

$$Coverage = \frac{1}{m}\sum_{i=1}^{m} max_{y \in Y_i} r_i(\lambda) - 1 \tag{4.37}$$

where m is the number of the instances in the evaluated data set; $r_i(\lambda)$ represents the rank of a label λ predicted by the ranking method; Y_i is the set of actual labels of a given instance x_i.

Ranking Loss expresses how many times the relevant labels are ordered in reverse. The performance is perfect when the ranking loss equals to zero. It is defined in Equation (4.38).

$$Ranking - Loss = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i||\bar{Y_i}|}|\{(\lambda_a, \lambda_b) : r_i(\lambda_a) > r_i(\lambda_b), (\lambda_a, \lambda_b) \in Y_i \times \bar{Y_i}\}| \tag{4.38}$$

where m is the number of the instances in the evaluated data set; λ_a and λ_b indicate class labels; $r_i(\lambda)$ represents the rank of a label λ predicted by the ranking method; Y_i is the set of actual labels of a given instance x_i and $\bar{Y_i}$ denotes the complementary set of Y_i in Y.

Average Precision (Equation (4.39)) measures the average fraction of labels ranked above a particular label $\lambda \in Y_i$, which actually are in Y_i. It derives from the performance measure used in information retrieval systems [152]. It is used to evaluate the document ranking performance for query retrieval. When average precision equals one, it means that the ranking runs perfectly.

$$Average - Precision = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i|}\sum_{\lambda \in Y_i}\frac{|\{\lambda^{'} \in Y_i : r_i(\lambda^{'} \leq r_i(\lambda))\}|}{r_i(\lambda)} \tag{4.39}$$

where m is the number of the instances in the evaluated data set; Y_i is the set of actual labels of a given instance x_i; $r_i(\lambda)$ represents the rank of a label λ predicted by the ranking method; $\lambda_{\prime}$ indicates the class label where the rank is higher than that of class label λ.

4.5.5 Multi-label Classification of SFRB

To evaluate the multi-label classification approaches for analyzing the multiple functions of SFRB, two data sets are used here: 372 SFRB located in central Scotland and 202 SFRB situated in Southern Baden. For each method and all data, during the training phase, one-leave-out cross-validation is performed. In comparison to the traditional method of splitting data into only two parts (i.e. training data and test data), cross validation is performed using different partitions on multiple rounds, and the validation results are averaged over the rounds. Therefore, in this study, to reduce the impact of variability of the

data splitting for classification, one-leave-out cross-validation has been applied. Specifically, each data instance from the data set is selected as the test data, and then the remaining data instances are used as the training data. In addition, five different values for each parameter were selected in a heuristic way suggested by the algorithms.

For comparison, the corresponding traditional classifiers SVM, KNN and BP are also applied in experiments. Multi-label classifiers MLSVM, MLKNN and BP-MLL were implemented in Matlab. The corresponding traditional classifiers used in this paper are available in WEKA (http://www.cs.waikato.ac.nz/ml/weka/). All experiments have been performed on a workstation with 2.4 GHz CPU and 2.0 GB RAM.

4.6 Spatial Analysis

4.6.1 Rational

Previous studies [153, 32] have indicated that the actual flood risk and the adaption measures are highly related with the characteristics of and developments within the area concerned. More recently, the Flood Directive suggests that flood risk management should be incorporated into spatial planning policies at all levels to enhance certainty and clarity in the overall planning process [53]. However, the spatial properties of these SFRB such as the water storage (which relates to flooding depth) in different regions of Scotland are ambiguous. Geo-statistics such as ordinary kriging, which is a spatial interpolation method, could provide a promising way to estimate numerical values for all key flood control variables everywhere in the study area. Moreover, the probability that certain threshold values are exceeded can also be calculated by using disjunctive kriging. Therefore,

an effective tool for SFRB in assessing spatial flood control is greatly needed and would be of great value for the effective flood risk managers.

4.6.2 Related Work

Kriging is a group of geo-statistical techniques designed to interpolate the value of a random field at an unobserved location derived from studies of its value at nearby sites. The variogram, which is a graph describing the spatial variation of a regionalized variable, is crucial in all kriging techniques [145]. It can be used not only for estimation but also to control a multivariate classification that the resulting classification are not too fragmented spatially to manage [125]. Numerous kriging techniques have been proposed, such as simple kriging, ordinary kriging, poisson kring and disjunctive kring. Different techniques are suitable for different practical purposes. For example, disjunctive kriging is used for non-linear problems and enables the error associated with a judgment to be converted to an estimated probability that the true value exceeds this threshold. Thereby it gives the decision-makers and designers a practical tool to judge the risk of taking the estimate at its face value. Various kriging techniques have been successfully applied in hydrogeology and remote sensing [188], natural resource management [126], environmental science [145] and agriculture [105]. More recently, geo-statistics has been applied in the wider area of water resource management, e.g. floodplain soil property mapping [50], river water quality monitoring [86] and river network discharge mapping [154]. An effective and rapid geostatistical tool providing examples of flood management involving SFRB, which are characterized by clear and relevant characterization variables, would support planning and communication among practitioners and planners.

4.6.3 Kriging Algorithm

To analyze the spatial statistics of given data with Kriging, the first step is variogram analysis. In geo-statistics, the variogram is a function describing the degree of spatial statistical dependence of a spatial random field (called spatial autocorrelation). A variogram interpolates a raster from a set of points using kriging (see below) with a known semivariogram model and its parameters [145]. The experimental variogram can be computed by using Equation (4.40).

$$\gamma(h) = \frac{1}{2n}\sum_{i=1}^{n}\{z(x_i) - z(x_i + h)\}^2 \tag{4.40}$$

where $\gamma(h)$ is the sample semi-variance at lag h, which is a vector in both distance and direction, $z(x_i)$ and $z(x_i + h)$ are the values of $Z(x)$ at locations x and $x + h$, respectively, and n is the number of pairs of comparisons separated by a lag h for $i = 1, 2, \cdots, n$.

The parameters used to describe variograms are nugget, partial sill and range. The parameter nugget is the height of the jump of the semi-variogram at the discontinuity at the origin. Sill is the limit of the variogram tending to infinity lag distance, whilst the range is the distance in which the difference of the variogram from the sill becomes negligible. And a partial sill is the sill minus the nugget. The empirical variogram, which provides information on the spatial autocorrelation of the SFRB dataset, is fitted by a mathematical function, which describes the structure of variation and ensures validity of the variogram. Before the production of any kriged map, the most suitable model is fitted by applying the least squares method to the points forming the empirical semi-variogram.

After variogram analysis, various krigning algorithms can be used. In this study, the most two popular kriging algorithms are introduced.

Ordinary Kriging: Ordinary kriging forms weights from a semi-variogram based on surrounding measured values to predict figures for unmeasured sites. The measured values nearest to the unmeasured sites have the greatest influence. Predictions were made for each location in the wider central Scotland area based on the semi-variogram and the spatial arrangement of measured values that are located within the vicinity.

Ordinary kriging provides best linear unbiased estimations with minimum error variance and is the most commonly used type of kriging. Furthermore, kriging weighs the surrounding measured values to derive a prediction for an unmeasured location. The general formula is formed as a weighted sum of the data (Equation (4.41)). The weight λ_i depends on a fitted model to the measured points, the distance to the prediction location, and the spatial relationships among the measured values around the prediction location.

$$Z(S_0) = \sum_{i=1}^{N} \lambda_i Z(S_i) \tag{4.41}$$

where $Z(S_i)$ denotes the measured value at the i^{th} location, λ_i is an unknown weight for the measured value at the i^{th} location, S_0 is the prediction location, and N means the number of measured values.

Disjunctive Kriging: Disjunctive kriging is a nonlinear generalization of kriging. This estimation technique allows for the conditional probability that the value of a spatially variable SFRB characterization parameter is greater than a cut-off level yet to be calculated [145, 105]. The method can be used in flood risk management decision-making to help determine when some action, such as the construction of a new SFRB or a change in reservoir management, is necessary. Two input parameters are required to use the technique: a cut-off level and the critical probability level [126]. The relevant standard formulae for disjunctive

kriging are shown in Equation (4.42) to Equation (4.44). Sample values x in the original space A are transformed into y in a normal space B that has a standard normal distribution such that $x = \varphi(y)$. The function φ is written as a linear combination of Hermite polynomials as shown in Equation (4.42) and discussed by Grad [55].

$$\varphi(y) = \sum_{k=0}^{\infty} C_k H_k(y) \tag{4.42}$$

where $\varphi(y)$ is the function of a linear combination of Hermite polynomials, C_k are coefficients to be calculated from the sample values $x_i (i = 1, 2, \cdots, n)$, and $H_k(y)$ is a Hermite polynomial of the order k.

In the space B, the value y has the so-called coordinates $H_0(y), H_1(y), \cdots, H_k(y), \cdots$ while in the space A, the corresponding value $x = \varphi(y)$ has the coordinates $C_0 H_0(y), C_1 H_1(y), \cdots, C_k H_k(y) \cdots$. Each sample value x_i is transformed to a value $y_i (i = 1, 2, \cdots, n)$ with coordinates $H_0(y_i), H_1(y_i), \cdots, H_k(y_i) \cdots$. To obtain the disjunctive kriging estimate μ_D^*, which belongs to the space A, someone first calculates the corresponding value ν_D^* in the space B. If $H_{0v}^*, H_{1v}^*, \cdots, H_{kv}^*, \cdots$ are the coordinates of v_D^* in B, these coordinates are obtained by a linear combination of the corresponding coordinates of the sample values. The estimator H_{kv} is of the form shown in Equation (4.43). To obtain μ_D^* from ν_D^*, the function φ is used as shown in Equation (4.44).

$$H_{kv} = \sum_{i=1}^{n} b_i H_k(y_i) \tag{4.43}$$

where H_{kv} is the coordinate of the normalized block estimator v_D in space B, calculated from the surrounding sample values $y_i (i = 1, 2, \cdots, n)$, and b_i are

weights, optimally defined through the use of the autocorrelation function of the sample values.

$$\mu_D = \varphi(v_D) = \sum_{k=0}^{\infty} C_k H_{kv} \tag{4.44}$$

where μ_D represents the disjunctive kriging estimator, and v_D is the corresponding value in normalized space B.

4.6.4 Spatial Distribution of SFRB

In this study, research is focused on flood-related variables such as *Engineered* (relative value, expressed in %, based on expert judgement of how engineered a basin appears in contrast to a natural water body), *Mean Flooding Depth* (average depth within the basin during flooding), *Maximum Flood Water Volume* (maximum volume of water within a basin during a typical flood) as discussed by Section 3.3, and two compound variables: *Managed Mean Flooding Depth* and *Managed Maximum Flood Water Volume*. The first compound variable can be derived from the variables *Mean Flooding Depth* and *Mean Depth of Basin*. *Managed Mean Flooding Depth* data for drinking water reservoirs are similar to the values collected for *Mean Flooding Depth*, while values for lakes are calculated by subtracting *Mean Depth of Basin* from *Mean Flooding Depth*. The second compound variable *Managed Maximum Flood Water Volume* is derived by multiplying *Managed Mean Flooding Depth* by *Flood Water Surface Area*. For details on how to define and determine the values for the variables, readers may refer to Chapter 3. All geo-statistical analysis were performed within the ArcGIS 9.2 [105].

4.7 Dam Failure Assessment

4.7.1 Rational

Historically, the underlying philosophy for dam safety has been stated as follows:

> "*the safety of a dam manifests itself in being free of any conditions and developments that could lead to its deterioration or destruction. The margins which separate the actual condition of a dam, or the conditions it is designed for, from those leading to its damage or destruction is a measure of its safety.*" [73]

Traditionally, dams are considered safe, because they have been built according to high technical standards. However, many dams that were constructed decades ago do not meet the current state-of-the-art dam design guidelines anymore. Moreover, many SFRB including reservoirs in Scotland and elsewhere are located immediately upstream of or adjacent to heavily populated areas, which will result in a greater hazard associate with dam failure. Due to an increase in awareness of these phenomena and climate change, people realize that all man-made structures have a potential risk of failure that has to be evaluated, assessed and managed [87]. For instance, flooding from reservoirs can result from an uncontrolled breach of the dam or overtopping during a severe rainfall. This may result in catastrophic consequences for life, property, critical infrastructure and economy [42, 162]. Therefore, risk associated with dams failure should be assessed. However, Bulletin 59 [73] has clearly recognized that dam failure is a complex process which includes engineered factors, human error in design, construction, operation, maintenance and surveillance. Moreover, detailed information for these factors are often costly,

time consuming or even unavailable. In this case, therefore, a rapid and cost-effective tool based on expert judgement will provide a promising way to assess dam failure.

4.7.2 Related Work

Various assessment tools have been studied for flooding risk associated with dam failures. For instance, the traditional "downstream hazard" assessment was used for dam safety evaluations, however, there existed internal inconsistencies in the approach [15]. More tools such as portfolio risk assessment [16], risk-based profiling system and condition indexing method [66], risk and reservoir in the UK [70], and dam safety management in Germany [140, 144] have been developed and applied worldwide. However, these methods are only applicable for initial risk assessment and should not be used alone for detailed dam failure assessment. Pitt [131] recommended the creation of inundation maps particularly for those areas near reservoirs. This recommendation has been implemented for most areas of the country [42, 166]. However, these maps are based on digital terrain models, and do not directly relate to any specific dam condition and associated risk of failure. They are merely one of many tools that help to assess risk. Keef et al. [88] adopted a model-based approach, which accounts for missing values. However, as the complexity of flooding risk increases, the number of missing data or even missing variables increases rapidly as well. Generally, critical information of interest for emergency services includes flood extent, water depth, flood water velocity, hazard level, time of initial inundation and time of peak arrival. However, the determination of these variables is relatively costly, time consuming and not very accurate. Moreover, the total complexity and process dynamic of freak storms can never be fully captured, and changes rapidly over time and in space [162].

4.7.3 Dam Failure Assessment Tool for SFRB

This study proposes a rapid tool for dam failure assessment of SFRB, taking the various factors mentioned in Bulletin 59 into account. Firstly, three new risk-related variables and their corresponding components are proposed by a group of international experts in engineering and science. Specifically, these experts include landscape planners, scientists and engineers from the University of Edinburgh, University of Freiburg, University of Munich and the SAWA project. After that, based on empirical study and international discussion, different weights are assigned to different components. Finally, each component (see Table 4.1 to Table 4.3) is split up into five bins, and guidance on how to determine values for each component is provided. (also see Section 3.3.41 to Section 3.3.43). Briefly speaking, the rapid tool is using Table 4.1 to Table 4.3 to survey and assess the dam safety of each SFRB. Here, it is important to distinguish clearly between hazard and risk. The hazard associated with a dam considers only the possible consequences of the structure failing regardless of their likelihoods. The risk of that failure occurring is determined by factors such as the lack of dam management and maintenance, potentially increasing the probability of failure. However, no statistical relationships can be used to accurately determine these complex variables, and expert judgement is therefore needed. The elements of the survey tool is introduced as follows.

Table 4.1: Brief characterisation of *Dam Condition* (%) in terms of five assessment bins, and its components and corresponding weightings.

Component	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
Dam Structure (30)	Very good condition; very tidy; most likely be well looked after by a water authority or council (>80%)	Good condition; most likely be looked after by a local authority or private owner (>60 to 80%)	Not tidy; potentially only little maintenance; minor signs of neglect, decay and erosion (40 to 60%)	Not tidy; most likely no maintenance since a long time; clear signs of decay and erosion (20 to <40%)	Very poor condition; surface cracking likely; seepages are present; dam unlikely to be still in use (<20%)
Spillway Condition (30%)	Usually concrete slipways; re-enforced grass-lined dams possible; no unwanted vegetation growth; well maintained (>80%)	Usually masonry or re-enforced grass-lined dams; no unwanted vegetation growth; well maintained; often well-integrated into the outlet (>60 to 80%)	Grass-lined dams; partially covered by unwanted vegetation; often well integrated within the outlet as a passive structure (25 to 60%)	Often a grass-lined earth dam with unruly vegetation cover; high proportion of vegetation growing within the pointing (5 to <25%)	No or at least no obvious spillway exists; obviously missing blocks; structure is in a state of urgent need of attention (<5%)
Wave Wall Condition (20%)	Well-maintained wave wall; concrete, masonry or large bolder; no unwanted vegetation growth (>90%)	Masonry or bolder; little proportion of pointing (i.e. mortar or cement between stones) containing vegetation (>70 to 90%)	Often a re-enforced earth dam; about 50% of pointing (i.e. mortar or cement between stones) containing vegetation (40 to 70%)	Often an earth dam; almost all the pointings (i.e. mortar or cement between stones) contain unwanted vegetation (10 to <40%)	Minor signs of erosion of the earth dam; clear indication of no maintenance; obvious missing blocks or masonry (<10%)
Operational Volume Impact (10%)	Operational volume often >20 km^3; major impact on dam due to constantly high pressure (>80%)	Operational volume often >10 and 20 km^3; major impact on dam due to high pressure (>60 to 80%)	Operational volume often >5 and 10 km^3; major impact on dam due to occasionally high pressure (40 to 60%)	Operational volume often >0.5 and 5 km^3; occasionally filled with water (20 to <40%)	Operational volume often 0.5 km^3; very rarely filled with water and therefore little pressure on dam (<20%)

Continued on next page

Table 4.1: (continued)

Component	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
Other Factors (10%)	Reservoirs currently operated by a water authority for drinking water or flood protection purposes (>85%)	Water bodies operated by councils, fishing groups and sailing clubs (>65 to 85%)	Partially managed water bodies; often warning information or other related signs indicating a legal responsibility waiver (25 to 65%)	Large semi-natural lake with little need for regular maintenance; sometimes an off-line retention basin (10 to <25%)	Small semi-natural lake, wetland or off-line water body without any need for regular maintenance (<10%)

Table 4.2: Brief characterisation of *Dam Failure Hazard* (%) in terms of five assessment bins, and its components and corresponding weightings.

Component	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
Overall Force on Dam (30%)	Dam height is often more than 35 m; very high overall force on dam; very poor construction regarding clay core, embankment or plastic liner; serious wave damage possible; blockage of spillway trash screen likely (>80%)	Dam height is often more than 30 m; high overall force on dam; poor construction of clay core, embankment or plastic liner; wave damage possible (>60 to 80%)	Dam height is often more than 25 m; sometimes high overall force on dam; some dam construction weaknesses (40 to 60%)	Dam height is often more than 18 m; rarely force on dam; good construction; good construction (20 to <40%)	Dam height is often less than 18 m; very rarely force on dam (empty basin); very good construction (<20%)
Potential Loss of Life (35%)	Very large water body; very high permanent and dense urban area proportion in the catchment just below of the failed dam; people are likely to be killed (>80%)	Large water body; high urban area proportion in the catchment just below of the failed dam; people may be killed (>60 to 80%)	Medium-sized water body; occasionally high urban area proportion in the catchment, but far away from the failed dam (40 to 60%)	Small water body; very low urban area proportion in the catchment, but just below of the failed dam (10 to <40%)	Very small water body; significant rural area; only few people can ever be effected in the remote area (<10%)

Continued on next page

Table 4.2: (continued)

Component	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
Importance of Infrastructure Affected by Dam Failure (25%)	Main airports, major railways, major roads and bridges, gas pipelines, key industry areas and key electricity supply structures are affected in a large urban area just below the failed dam; very serious loss of water (>80%)	Railways, major roads, bridges, major industry and electricity supply structures are affected in a small urban area just below the failed dam; very serious loss of water (>60 to 80%)	Railways, roads, schools, small industry and electricity supply structures are affected in an urban area located faraway from the failed dam (40 to 60%)	Farming infrastructure, minor roads and electricity supply structures are affected in an urban area faraway from the failed dam (5 to <40%)	Significantly high rural area proportion; no important infrastructure; only single-track roads affected in a remote area (<5%)
Other Factors Influencing Dam Failure Hazard (10%)	Serious erosion and damage of the dam; cloudy seepage or leakage water; dam covered by trees and bushes (hazard of internal erosion); blocked or damaged spillway (hazard of overtopping); severe rodent and/or crayfish attack; key nature protection area below dam (>80%)	Erosion and damage of the dam; some trees and bushes; slightly damaged spillway; minor rodent attack; nature protection area below dam (>50 to 80%)	Little erosion and damage of the dam; some bushes and high grass; some reeds at wet areas; apparent reasons to raise concern (25 to 50%)	All dam structures are well-maintained; clear spillway and pipes; short grass cover; no indication of any imminent hazard (10 to <25%)	Semi-natural water body with no apparent signs of any anticipated hazard; public park (<10%)

Table 4.3: Brief characterisation of *Dam Failure Risk* (%) in terms of five assessment bins, and its components and corresponding weightings.

Component	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
Structural Failure Risk (20%)	Very high dam pressure; poorly maintained; unsafe operation; very narrow, tall and long dam with many potential weak points (>80%)	High dam pressure; poorly maintained; unsafe operation; very narrow and tall dam (>60 to 80%)	Medium dam pressure; poor maintenance; unsafe operation; narrow, tall and long dam (40 to 60%)	Medium dam pressure; good maintenance; safe operation; safe structures such as a wide and shallow dam (10 to <40%)	Low dam pressure; very good maintenance; safe structures such as a very wide, shallow and short dam (<10%)
Loss of Life Risk (50%)	Very dense urban areas with a very high proportion of permanent population in the catchment just below the failed very high dam; no apparent emergency plan in operation (>80%)	Dense urban areas with a high proportion of permanent population in the catchment just below the failed high dam; poor emergency plan in operation (>60 to 80%)	Dense urban areas with a high proportion of permanent population in the catchment far from the failed medium-sized dam; only some emergency planning evident (35 to 60%)	Sparse urban areas with a low proportion of permanent population in the catchment just below the failed small dam; minor emergency measures required (10 to <35%)	Sparse urban areas with a low proportion of permanent population far from the failed very small dam; no need for any emergency planning (<10%)
Risk of Infrastructure Failure (20%)	Poorly protected from flooding; located in low lands; low embankments; close to watercourse; very deep flood water; no adaptation to climate change evident (>80%)	Poorly protected from flooding; located in low lands; high embankments; close to watercourse; deep flood water; little evidence for adaptation to climate change (>60 to 80%)	Poorly protected from flooding; located in high lands; high embankments and close to watercourse; shallow water; some adaptation to climate change apparent (40 to 60%)	Well protected from flooding; located in high lands; low embankments; far from watercourse; shallow water; adapted to climate change (15 to <40%)	Well protected from flooding; located in high lands, high embankments; far from watercourse; very shallow water; fully adapted to climate change (<15%)

Continued on next page

Table 4.3: (continued)

Component	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
Other Factors Influencing Dam Failure Risk (10%)	Excessive embankment erosion during particularly wet years likely; risk of contaminated sediment deposition in populated areas due to prolonged flooding; very likely failure of a further reservoir downstream; extreme shifts in weather patterns due to climate change likely; possibility of reservoir cascade failure; risk of war damage; risk of terror attack or sabotage; seismic activity likely (>80%)	Excessive embankment erosion during particularly wet years likely; risk of contaminated sediment deposition in populated areas due to prolonged flooding; likely failure of a further reservoir downstream; extreme shifts in weather patterns due to climate change likely; occasional seismic activity (>60 to 80%)	Embankment erosion during particularly wet years likely; shifts in weather patterns due to climate change likely; potential failure of a further reservoir located downstream (30 to 60%)	Embankment erosion during particularly wet years likely; shifts in weather patterns due to climate change likely (10 to <30%)	Well maintained; safe operation; no apparent risk identified (<10%)

In the context of dam failure risk and hazard, the sites of interest are only those, which have a dam and may be able to play a role in flood management. Structures that have a flood control function are considered to be those where the water level can be controlled either manually or automatically, and are typically former or current engineered water supply reservoirs in Scotland. Therefore, among the data set, precisely 199 sites are selected for the dam failure assessment of SFRB in central Scotland (Figure 4.5). Based on these selected basins, the dam failure risks

and hazards for different types of SFRB are analyzed and the spatial distribution of dam failure hazards and risks across the study area are mapped.

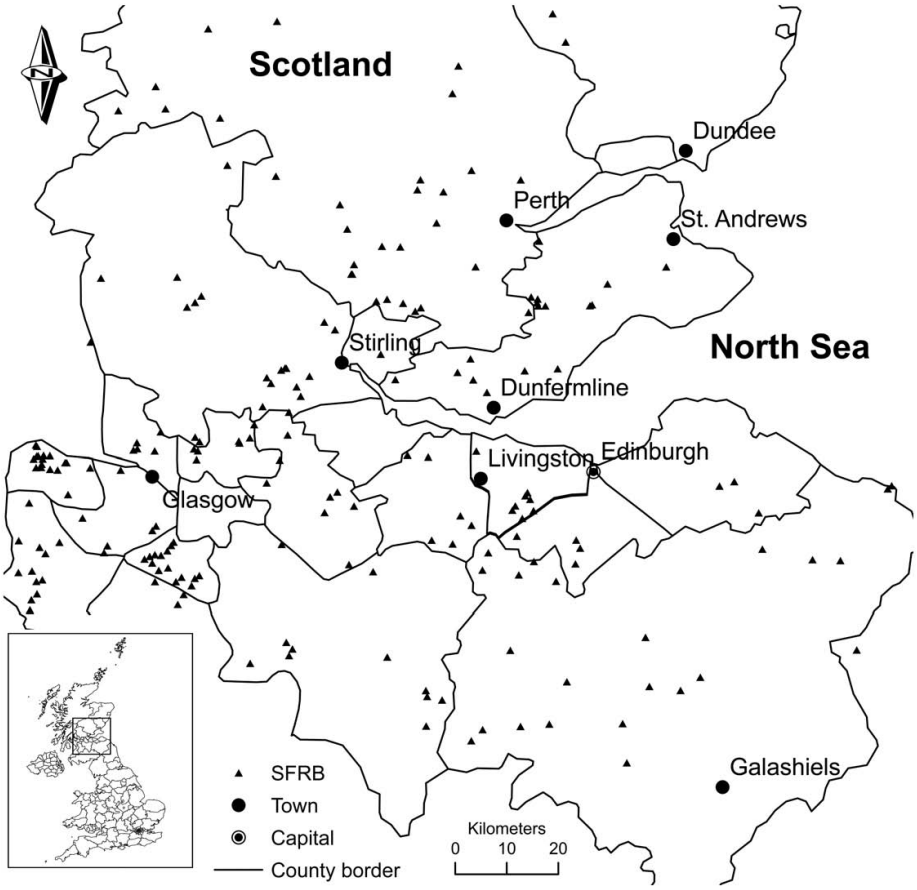


Figure 4.5: Study area, administrative boundaries and the 199 sustainable flood retention basins (SFRB) with dams in central Scotland area (United Kingdom).

4.8 Summary

As a whole, this chapter presented the methodology used in the thesis. To better explore and assess SFRB, various methods were applied to reveal the hidden patterns of the SFRB data. Specifically, these methods mainly include hierarchial clustering, PCA, SOM, feature selection, multi-label classification, Kriging, etc.

For each method, its rationale, related work and theory were elaborated. To verify the performances of these methods, experiments were thus undertaken on the SFRB data. The screening tool for dam failure assessment of SFRB was finally described.

Chapter 5

RESULTS AND DISCUSSION

This chapter presents the results of various methods applied in the thesis (stated in Chapter 4). The findings of each method are demonstrated and discussed in detail. More specifically, Section 5.1 illustrates the cluster analysis of Scottish SFRB. Section 5.2 exhibits the results of Principal Component Analysis (PCA). Section 5.3 demonstrates the application of Self-organizing Map(SOM) on SFRB, part of which has been published in the Proceedings of the 12th International Water Association International Conference on Wetland Systems for Water Pollution Control. Section 5.4 presents the variable selection and the performance of classification based on the selected variables, which has been published on the journal of Water Research. Exploration of multi-label classification and the application on representative case studies are depicted in Section 5.5. Geo-statistics of SFRB regarding flood control is analyzed in Section 5.6 and published on Water and Environment Journal. The dam failure assessment of SFRB is presented in Section 5.7. Finally, Section 5.8 summarizes this chapter.

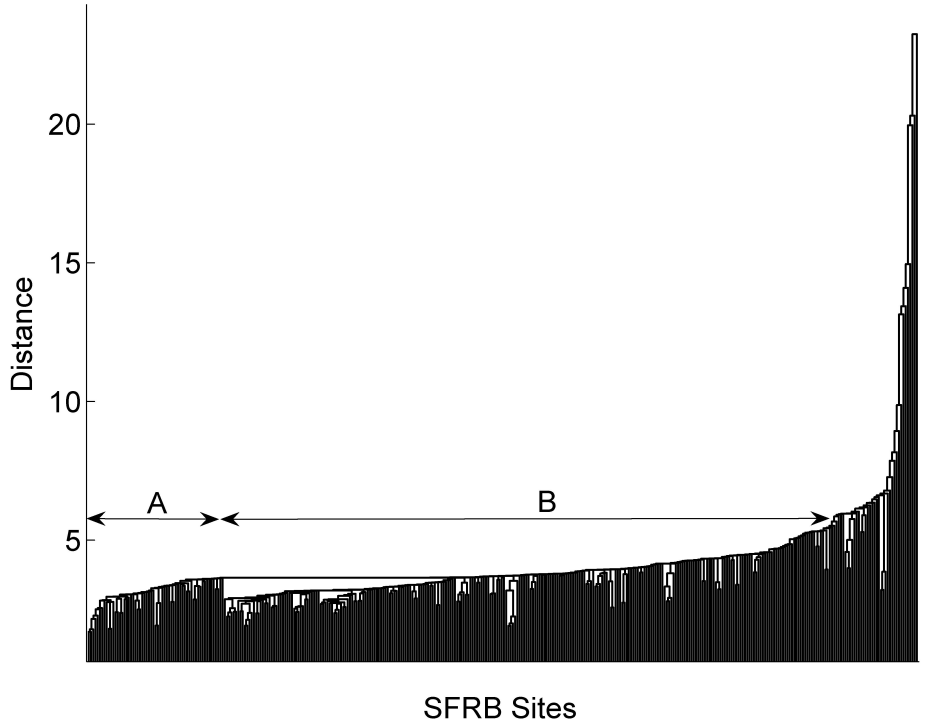


Figure 5.1: Dendrogram based on 43 variables for the data set of 372 retention basins with Single link and Euclidian distance used to identify Sustainable Flood Retention Basin types.

5.1 Clustering of SFRB

To reveal the underlying pattern of Scottish SFRB data, the agglomerative hierarchical clustering was applied (cf. Section 4.1.3). The goal is to find the intrinsic clusters (objects in the same cluster have similar properties) in the SFRB data set, which could reasonably explain the types of SFRB. Here, the Euclidean distance, the most common distance measure, was used to calculate the pair-wise distances between observations. Different linkage criteria including single link, average link, complete link and ward link were adopted for clustering respectively (cf. Section 4.1.3). For each criterion, the clustering creates a hierarchical cluster tree, known as a two dimensional dendrogram, which illustrates the fusion of each successive stage of the analysis. Figure 5.1 shows the dendrogram of clustering

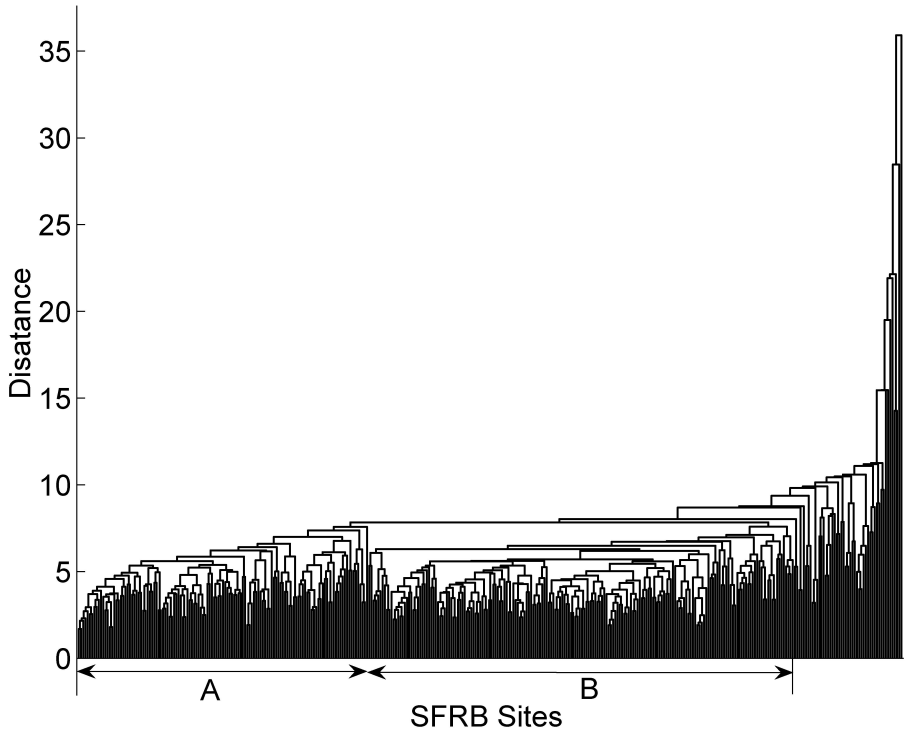


Figure 5.2: Dendrogram based on 43 variables for the data set of 372 retention basins with Average link and Euclidian distance used to identify Sustainable Flood Retention Basin types.

based on single linkage criteria. It is clear that only two clusters A and B could be identified. It is difficult to distinguish the finer clusters, thus it is hard to discover the intrinsic structure of the SFRB. Similarly, Figure 5.2 illustrates the hierarchical clustering tree based on average linkage, where only two clusters A and B are shown to be obvious. Since there is no significant difference of the distances between two clusters, the hierarchies are not distinct. Therefore, it is also a non-trivial task to clearly distinguish the clusters from the dendrogram, even though its performance is slightly better than that based on the single linkage (Figure 5.1).

Hierarchical clustering based on complete linkage was performed on the same dataset and a dendrogram as generated as shown in Figure 5.3. To find an

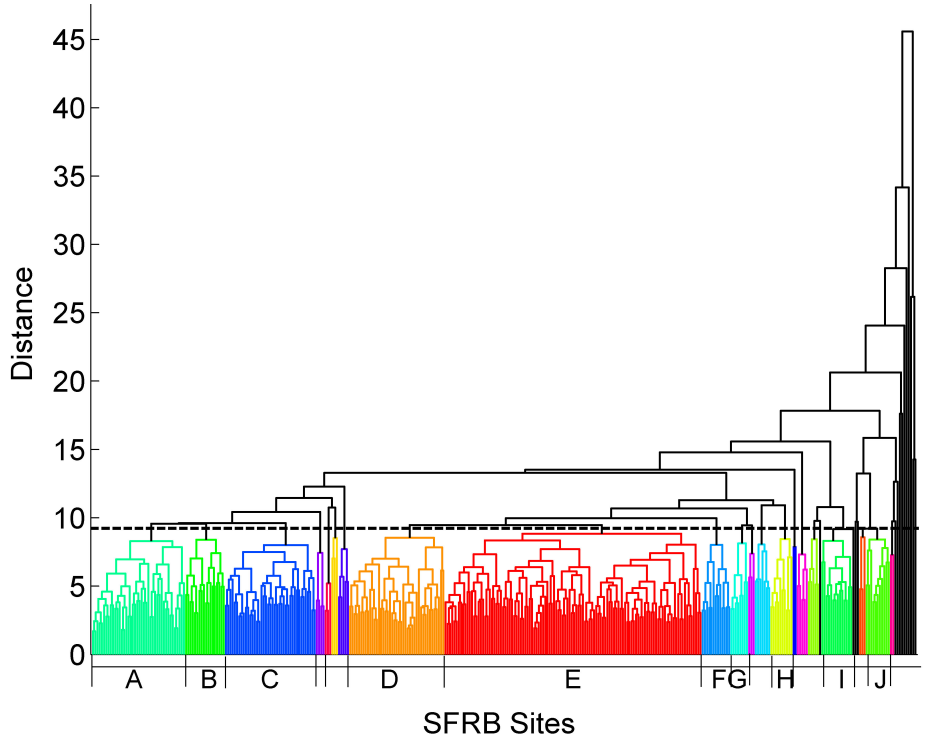


Figure 5.3: Dendrogram based on 43 variables for the data set of 372 retention basins with Complete link and Euclidian distance used to identify Sustainable Flood Retention Basin types.

appropriate number of clusters, a dashed line was drawn across the dendrogram graph and 35 nodes were formed. The figure demonstrates the whole Scottish SFRB data was clustered into 35 groups, where each color represents a group. The groups which include more than 7 SFRB were viewed as intrinsic clusters marked by A-J, while the groups which contain only a few objects (less than 7 SFRB) were regarded as outliers. In total, 330 SFRB were distributed in 10 intrinsic clusters (A-J from left to right) and 42 SFRB were identified as outliers. Compared with the ground truth of SFRB types, clusters A (42 sites), B (18 sites), C (41 sites) and J (14 sites) correspond to SFRB type 2 with precisions of 88.1%, 66.7%, 92.7% and 71.4%, respectively. Clusters E (116 sites), F (13 sites) and I (15 sites) represent SFRB type 6, which have precisions of 70.7%, 84.6% and 80%, respectivey. Cluster G (8 sites) and H (10 sites) contain SFRB of type

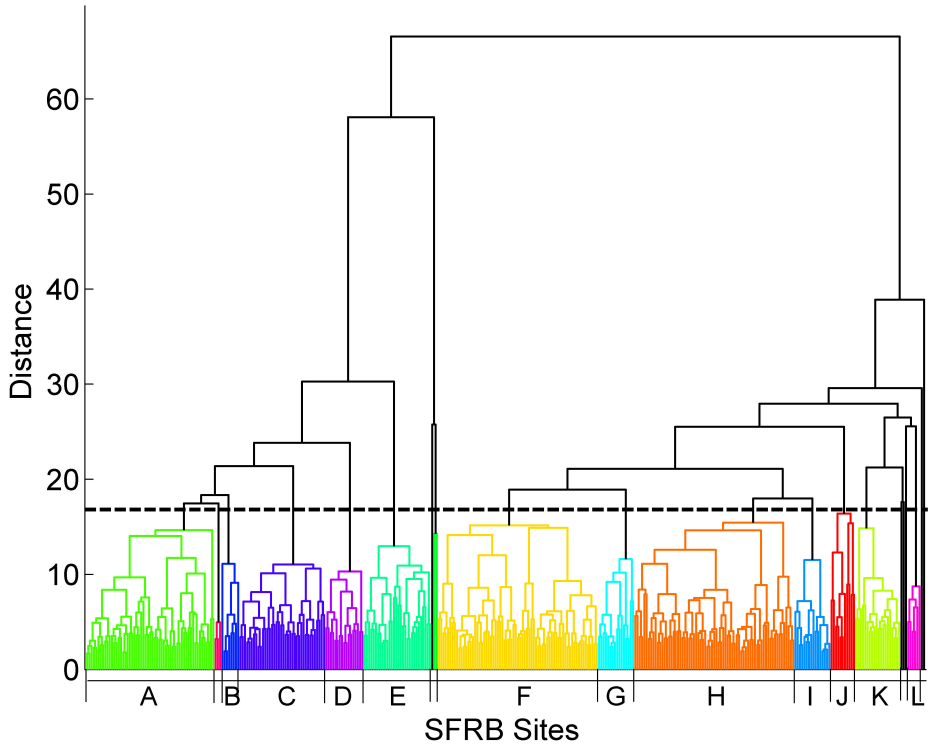


Figure 5.4: Dendrogram based on 43 variables for the data set of 372 retention basins with Ward's link and Euclidian distance used to identify Sustainable Flood Retention Basin types.

5, with relatively low precisions of 50% and 70% respectively. Particularly, cluster D is a mixture of five SFRB types, where types 2 and 6 are slightly dominant. Notably, there were no specific clusters matching with SFRB types 1, 3 and 4. In addition, when the distance axis is about 17 on Figure 5.3, the whole data set was clustered into two main groups: cluster J and another large group covering clusters A-I. This implies that the SFRB in cluster J (mainly type 2) differ significantly from others and the SFRB types 2, 5, and 6 were forced in one group. However, this is not reasonable, because in principle, SFRB in type 2 should be similar. Furthermore, generally the SFRB of type 2 were significantly different from SFRB of type 6 and thus should not be grouped together.

Similarly, based on the Ward criteria, clustering was applied on the Scottish SFRB

Table 5.1: Comparison of Ward's clustering results and the ground truth of SFRB types.

Cluster	#sites	T1	T2	T3	T4	T5	T6	CT	Precision	Recall
A	57	0	52	2	2	1	0	2	91.2	38.5
B	8	0	8	0	0	0	0	2	100	5.9
C	38	1	35	0	0	1	1	2	92.1	25.9
D	17	0	10	2	5	0	0	2, 4	58.8	7.4, 29.4
E	30	6	24	0	0	0	0	1, 2	80	66.7, 17.8
F	71	0	1	2	4	18	46	6	64.8	31.3
G	16	0	1	3	0	8	4	5	50	21.1
H	71	0	0	12	1	7	51	6	71.8	34.7
I	16	0	0	8	0	0	8	3,6	50	27.6, 5.4
J	11	1	1	0	0	0	9	6	81.8	6.1
K	20	0	0	0	1	3	16	6	80	10.9
L	6	0	0	0	0	0	6	6	100	4.1
Ground Truth		9	135	29	14	38	147			

Note: 372 sites in total (361 SFRB, 11 outliers); T (Type); CT (Corresponding Type).

data, resulting in the dendrogram as shown in Figure 5.4. The data were clustered into 20 groups with the dash line, where 12 groups (containing 361 SFRB) were intrinsic clusters assigned with A-L while the other 8 groups (11 sites), each of which only contains one or three SFRB, were regarded as outliers (non-SFRB). To evaluate the efficiency of the methodology, the Precision and Recall of each cluster was calculated by comparing the clusters with the ground-truth types of SFRB judged by experts (see Table 5.1). The Precision equals the number of SFRB which are correctly predicted divided by the number of SFRB in the cluster. Recall is defined as the number of SFRB which are correctly predicted as one type divided by the number of SFRB of ground truth in this type. The Table 5.1 shows that Clusters A (57 sites), B (8 sites) and C (38 sites) correspond

to SFRB type 2 achieving precisions of 91.2%, 100% and 92.2% respectively. In total, they contained 95 out of 135 SFRB of type 2 (ground truth). Most of SFRB included in clusters F (71 sites, precision 64.8%), H (70 sites, precision 71.8%), J (11 sites, precision 81.8%), K (20 sites, precision 80%) and L (6 sites, precision 100%) belonged to type 6. In total, 127 out of 147 SFRB (ground truth) in type 6 were included in these clusters. Cluster I (16 sites) corresponds to SFRB type 3 as well as type 6 for precision of 50%. In cluster D (17 sites), although type 2 accounts most of the cluster (10 out of 17 SFRB), the recall for type 4 (29.4%) is much higher than that of type 2 (7.4%). Therefore, cluster D represents both SFRB type 4 and type 2. Cluster E (30 sites) includes 24 SFRB with type 2 and only 6 SFRB with type 1, however, its recall for type 1 is as high as 66.7%. So cluster E can be viewed as SFRB type 1. It indicates that to some extent, the other 24 SFRB of type 2 have similar characteristics with the SFRB in type 1. Cluster G stands for SFRB type 5 with the precision of 50% with low recall of 21.1%, but type 5 has higher recall of 47.4% in cluster F.

Each cluster can be directly linked to a SFRB type. The mixed distribution of cluster entries in the corresponding SFRB types was explainable. Because in reality, one SFRB often have multiple functions and its main function might change over time. For example, some SFRB were originally built for drinking water supply purposes (belonging to type 2), however, after hundreds of years, the reservoirs were used as fishing ponds or constructed wetlands. It means that the reservoirs have 'shifted' from type 2 to types 4 or 5 which have more aesthetic or recreational functions.

In general, all six SFRB types were identified and demonstrated by clusters in Figure 5.4. The largest two groups were Natural Flood Retention Wetlands (Type 6, 179 sites) and Traditional Flood Retention Basins (Type 2, 133 sites). At the top of the figure, two large groups finally merged together, one of which were mainly SFRB type 2 and type 1 (cluster A-E) and one of which was comprised of

types 3, 4, 5, and 6 (cluster F-L). This is in line with the fact that these two large groups are significantly different from each other. In addition, the Ward's link clustering outperformed the other three hierarchical clustering methods, revealing the intrinsic structures of the SFRB data more efficiently and effectively.

5.2 Feature Reduction with PCA

Principal Component Analysis (cf. Section 4.2.3) was applied to project the Scottish SFRB data into a new feature vector, trying to use the first few (e.g. L) principal components to show the multivariate data in a lower-dimensional space. The question arose as to how to select a value for L. Each principal component (eigenvector) was associated with a weight (eigenvalue) obtained from the covariance matrix of SFRB data. The key point is to choose an appropriate value for L such that the cumulative weight of the first L principal components is above a particular threshold (90% in this study). As a result, the first L principal components would be used to represent the original data with little information lost. Table 5.2 displays the 43 principal components, the corresponding weights and the cumulative weights. The last column indicates that the cumulative weight of the first 23 principal components was 90.9%, which just achieved the threshold of 90%. Therefore, here L should be 23. This means that the original 43-dimensional SFRB data can be represented by the first 23 principal components.

However, each principal component is a linear combination of 43 original variables, so it is difficult to interpret the SFRB data in the new feature space by using the original variables. To find relationships between the principal components and the original variables, bi-plot was used. Bi-plot allows visualizing the magnitude and sign of each variable's contribution to each pair of principal components in a 2-dimensional figure. In the visualizing figure, the longer the line (from the

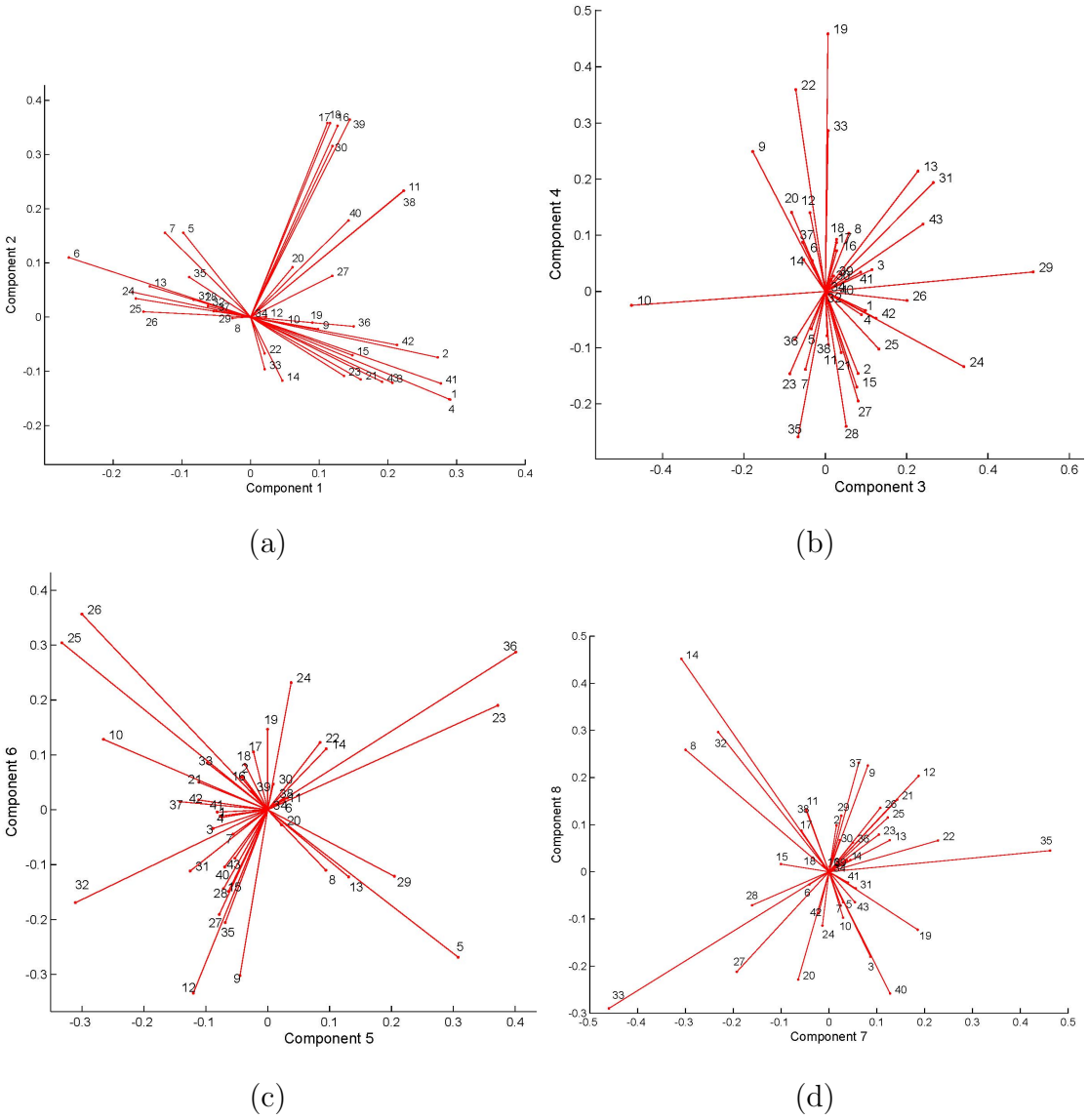


Figure 5.5: Visualization of the magnitude and sign of each variable's contribution to the (a): first and second components, (b): third and forth components, (c): fifth and sixth components, (d): seventh and eighth components.

origin to each destination) is, the more the corresponding variable contributes to the two components. Figure 5.5 gives four examples of bi-plot for the first 8 principal components. For example, Figure 5.5(a) displays the bi-plot for the first and the second principal components. It was clear that variables 6, 17, 18, 16, 39, 30, 11, 41, 1, 2, and 4 were important due to their high contributions

Table 5.2: Statistics of principal components.

C	W	CSW	CSWP(%)	C	W	CSW	CSWP(%)
1	8.76	8.76	20.9	23	0.53	38.18	90.9
2	5.83	14.59	34.7	24	0.50	38.68	92.1
3	2.53	17.12	40.8	25	0.44	39.13	93.2
4	2.18	19.30	45.9	26	0.43	39.55	94.2
5	1.77	21.07	50.2	27	0.41	39.96	95.1
6	1.69	22.76	54.2	28	0.37	40.33	96.0
7	1.44	24.20	57.6	29	0.32	40.65	96.8
8	1.40	25.59	60.9	30	0.29	40.94	97.5
9	1.27	26.87	64.0	31	0.24	41.18	98.0
10	1.15	28.01	66.7	32	0.19	41.37	98.5
11	1.10	29.11	69.3	33	0.18	41.55	98.9
12	1.00	30.11	71.7	34	0.15	41.70	99.3
13	0.92	31.04	73.9	35	0.13	41.83	99.6
14	0.91	31.95	76.1	36	0.08	41.91	99.8
15	0.87	32.82	78.1	37	0.04	41.96	99.9
16	0.84	33.66	80.1	38	0.02	41.98	99.9
17	0.82	34.48	82.1	39	0.01	41.99	100.0
18	0.75	35.23	83.9	40	0.00	42.00	100.0
19	0.67	35.91	85.5	41	0.00	42.00	100.0
20	0.64	36.54	87.0	42	0.00	42.00	100.0
21	0.56	37.11	88.3	43	0.00	42.00	100.0
22	0.54	37.65	89.6	Sum	42		

Note: C (Component); W (Weight); CSW (Cumulative Sum of Weight); CSWP (Cumulative Sum of Weight in Percent).

to the first two principal components. In contrast, Figure 5.5(b) demonstrates the bi-plot for the third and fourth principal components, where the related important variables were 19, 22, 29, 28, 35, and 10. For the fifth and sixth principal components, variables 25, 26, 36, 23, 5, 9, 12 and 32 (Figure 5.5(c))

Table 5.3: Each variable's contribution to PCA (in a descending order).

Order	Contribution	Variable ID	Order	Contribution	Variable ID
1	5.69	21	23	4.94	4
2	5.68	13	24	4.92	14
3	5.65	43	25	4.87	6
4	5.58	27	26	4.86	1
5	5.56	3	27	4.84	35
6	5.44	23	28	4.83	22
7	5.42	40	29	4.77	16
8	5.39	11	30	4.73	28
9	5.36	15	31	4.73	18
10	5.34	38	32	4.62	32
11	5.28	9	33	4.61	17
12	5.27	42	34	4.50	33
13	5.23	7	35	4.48	12
14	5.20	20	36	4.48	30
15	5.18	41	37	4.42	19
16	5.15	5	38	4.21	8
17	5.12	25	39	4.12	37
18	5.11	31	40	4.05	39
19	5.09	2	41	4.02	10
20	5.06	24	42	3.81	29
21	5.05	36	43	0.00	34
22	4.96	26			

were regarded as important contributors, while the variables 8, 32, 14, 35, 40 and 33 in Figure 5.5(d) contributed much to the seventh and eighth principal components. Similarly, important variables could be identified for the remaining pairs of dimensions.

Findings from Figure 5.5 implied that it was difficult to determine the important

variables for different components. Moreover, it was impossible to visualize all the variables in a 2-dimensional space and thus impossible to find the original variables which were important for all the components in a single figure. To quantitatively evaluate each original variable's contribution to all 23 principal components, the eigenvalues and eigenvectors were combined. The contributions are listed in descending order in Table 5.3. The higher the contribution is, the more the variable contributes to the 23 components in the new feature space. Table 5.3 shows that the highest contributor is *Impermeable Soil Proportion* (variable 21), which should be the most important variable for all components accordingly. The most important contributors were variables 13, 43, 27, 3, etc. Notably, the contribution of variable 34 (*Viniculture Catchment Proportion*) was zero, since there was no viniculture in the research area (central Scotland) and the variable was not applicable to Scottish SFRB. Moreover, the contributions of the variables were slightly different, ranging from 5.69 to 3.81, which indicates that all the variables (except for variable 34) are relatively important for the 23 principal components of the newly transformed data. Therefore, to some extent, PCA can help to reduce SFRB dimensions (from 43 to 23),however, it is not the best way to identify the key variables, .

5.3 SFRB Analysis with SOM

5.3.1 Assessment of the Relationships between Variables

The Self-organizing Map (SOM) model (cf. Section 4.3.3) was applied on SFRB data to identify the relationships between all 43 characteristic variables. Firstly, the SFRB data was normalized such that each variable has unit variance. After that, the data was used to train the SOM model and finally the feature map was created to visualize results (e.g. Figure 5.6). Its final quantization error

was 3.961 and the final topographic error was 0.022. Figure 5.6 displays the basic visualization of unified distance matrix (U-matrix) along with 43 component planes. Specifically, the U-matrix provide a visual representation of the distance between the neighbouring map neurons (calculated based on all the variables) and thus help to see the cluster structure of the map. High values on the U-matrix indicate cluster borders while the uniform areas of low values represent clusters. Figure 5.6 indicates that the SFRB data only contain three clusters, including one large cluster (in blue) and two small clusters (at the left bottom corner). Each of the component planes show the values of each variable of each map unit. The values of variables were denormalized to their original scales and indicated in color bars. Similar patterns between two variables indicated they had similar contribution to the SFRB system. Therefore one of the variables can be eliminated if necessary. For instance, variables 1, 4, 21, 41, and 43 represented similar patterns whereas the variable 6 showed an opposite pattern. This is in line with the fact that the highly *Engineered* SFRB usually have a good outlet arrangement and are well maintained and thus often have good *Dam Condition*. However, the higher engineered the structure was, the more difficult for the land animal to pass by. This finding also verified the correlations previously identified using PCA (Figure 5.5(a)). Another example is that the variables 11, 16, 17, 18, 30, 38, 39 and 40 had similar patterns with the U-matrix, which implied that these variables played key roles in grouping the SFRB. Furthermore, the pair-wise variables (2 vs. 42, 8 vs. 37, 25 vs. 26, and 24 vs. 29) also demonstrated similarity. In this context, the variables showing similar patterns can be replaced by only one of the variables. As a result, the redundant variables can be removed.

Notably, variables 10 and 12 did not demonstrate an important contribution to SFRB distinction due to their low variance. The findings were consistent with the results of PCA shown in Figure 5.5(a), where these variables were located near the origin which indicates their low contribution to the first two components.

Variable 34 can be ignored since this variable was not applicable to SFRB in central Scotland.

The application of the SOM model visualized the relations among the 43 variables, validated the findings of PCA analysis, and achieved feature reduction. The main advantage of the application of SOM is, however, the neural networks ability to learn the data structure and then predict variables and SFRB types (see below Section 5.3.2 and Section 5.3.3), which cannot be achieved by PCA and cluster analysis.

5.3.2 SFRB Variables Prediction

In theory, the SOM model can be applied to predict all the characteristic variables of SFRB. But it is only meaningful to predict the expensive-to-determine and time-consuming variables (target variables), e.g. *Mean Flood Depth*, *Mean Sediment Depth*, by using easy-to-determine, cost-effective and reliable variables such as *Dam Height*, *Annual Rainfall*. The selection of these easy-to-determine variables can either be based on expert judgment (specifying the variables in advance) or based on statistical analysis. Since expertise is often unavailable and expert judgment is subjective to some extent, the statistical method, which concerns the correlations among variables, is preferred. First, the correlation coefficients among all variables and their corresponding P-values were calculated. If P-value of any two variables was less than 0.01, it meant they had significant correlation with each other. Then all the related key variables, which have P-values <0.01were found for the variables to be predicted. Table 5.4 lists the interesting target variables and their corresponding highly related variables. For example, concerning variable 9, its highly related key variables were 1, 4, 6, 7, 12, 24, 25, and 41. Based on these key variables, the SOM model was applied to predict the target variables. After running the simulation, the predicted

SFRB variables were subsequently compared with the original real data. Their comparisons were visualized in sub graphs of Figure 5.7 separately. To evaluate the effectiveness of the prediction, the normalized root mean squared error (NRMSE) and P-values were chose as criteria (see Table 5.4). The smaller the NRMSE was, the better the performance was. If the P-value >0.05, it meant that the predictions had no significant difference with the actual data. Thus, the bigger the P-value, the closer the predictions were to the real data. The predictions of variables 1, 4, 16, 18, 41 had low NRMSE and high P-values, which indicated that SOM achieved good performance. In contrast, the predictions for variables 13 and 28 were poor. Since their NRMSE were high and the P-values were close to 0.05, which implied that there were large difference between the predictions and the real data.

Table 5.4: Prediction of SFRB variables and types by using SOM model.

ID	NRMSE	P-value	Variables used for prediction
1	0.32	0.94	2 3 4 5 6 7 9 11 13 15 19 21 23 24 25 26 27 28 35 36 38 41 42 43
4	0.32	0.91	1 2 3 5 6 7 9 11 13 15 19 21 23 24 25 26 27 31 35 36 38 41 42 43
6	0.57	0.74	1 2 3 4 5 7 9 11 13 15 21 23 24 25 26 27 31 35 36 38 41 42 43
7	0.96	0.64	1 2 3 4 5 6 9 13 23 24 32 35 36 41 42 43
9	0.64	0.64	1 4 6 7 12 24 25 41
11	0.92	0.16	1 2 3 4 6 13 15 16 17 18 20 21 24 25 26 27 30 36 38 39 40 41 42 43
13	1.37	0.07	1 2 3 4 6 7 11 15 21 23 25 26 27 38 41 42
15	0.87	0.79	1 2 3 4 6 11 13 21 27 38 41 42 43
16	0.17	0.97	11 17 18 19 27 30 38 39 40
17	0.48	0.65	11 16 18 19 27 30 38 39 40
18	0.23	0.72	11 16 17 19 27 30 38 39 40
20	0.97	0.10	11 22 27 38 40
21	0.83	0.30	1 2 3 4 5 6 11 13 15 23 38 41 42 43
24	0.80	0.96	1 2 3 4 6 7 9 10 11 12 19 25 26 38 41 42 43
25	0.79	0.73	1 2 3 4 6 9 11 13 23 24 26 29 35 36 38 40 41 42 43
26	0.64	0.57	1 2 3 4 6 11 13 23 24 25 27 29 38 40 41 42 43
27	0.87	0.10	1 2 3 4 6 11 13 15 16 17 18 20 22 26 28 38 39 40 41 42
28	1.25	0.25	1 19 22 27 41 43
30	0.51	0.58	11 16 17 18 38 39 40
38	0.97	0.13	1 2 3 4 6 11 13 15 16 17 18 19 20 21 24 25 26 27 30 36 39 40 41 42 43
41	0.42	0.71	1 2 3 4 5 6 7 9 11 13 14 15 19 21 23 24 25 26 27 28 31 35 36 38 42 43
42	0.85	0.97	1 2 3 4 5 6 7 11 13 15 21 24 25 26 27 35 36 38 41 43

Continued on next page

Table 5.4: (continued)

ID	NRMSE	P-value	Variables used for prediction
43	0.80	0.97	1 2 3 4 5 6 7 11 15 19 21 23 24 25 26 28 35 36 38 41 42
Type	0.42	0.91	1 2 3 4 5 6 7 9 11 13 15 19 21 23 24 25 26 27 28 35 36 38 41 42 43

To demonstrate the performance of the SOM model on the prediction of SFRB variables, the differences between the actual and the predicted values for the interesting target variables were illustrated on a series of figures 5.7. For instance, Figure 5.7(a) shows the residual plot for the variable of *Engineered*. In combination with figure 5.8, it indicated that the big differences ($\pm$ 25%) arose in SFRB of types 1, 3, 4 and 5, while the differences for SFRB of types 2 and 6 were relatively small. Specifically, the SFRB in type 1 were underestimated, it might be because they have many common characteristics with SFRB of type 2 and thus were treated as type 2. In fact, they are mainly used to feed hydropower stations and usually have higher engineered structures and thus receive high values for *Engineered* in practice. The predictions for types 3, 4 and 5 did not match well with the observed values, the reason might be that the samples of types 3, 4 and 5 were insufficient during the SOM model training. Many SFRB of type 6 were overestimated, which means that they were predicted more 'engineered' than the actual observations. Figure 5.7(b) shows the residual plot for *Land Animal Passage*. It indicated that most of the differences between the predictions and the actual values varied in the range of $\pm$ 15%. The predictions of *Mean Flooding Depth* and *Maximum Flood Water Volume* (Figure 5.7(c) and (d)) fitted with the actual data well except for several particular basins which have extreme deep water (e.g. loch Lomond). Regarding the *Dam Failure Hazard* and *Dam Failure Risk* (Figure 5.7(e) and (f)), there were no big differences between the actual and the predicted values except for very few exceptions where the poor dam conditions were observed on site.

5.3.3 SFRB Types Prediction

Treating the type of SFRB as a variable,the SOM model can also be applied to predict the types of SFRB in a similar procedure. Based on the correlation efficiencies between the 43 variables and the SFRB types, 25 variables (see Table 5.4) which have high correlations with the target were selected to predict the types of SFRB. Figure 5.8 displays the predicted results and the real types of SFRB. It is clear that the predictions for type 2 and 6 performed well. For details, Table 5.5 illustrates the comparison by showing the distribution of the predicted types of SFRB. For type 2 (containing 67 SFRB in real), 60 SFRB were predicted correctly by using SOM model based on 25 related variables, while the other 7 SFRB were estimated as type 3. Concerning the 73 SFRB which actually belonged to type 6, 46 sites were predicted as type 6 and 25 sites were predicted as type 5. In particular, types 1, 3, 4, and 5 were poorly predicted. For instance, all 4 SFRB in type 1 were estimated as type 2. The predictions for types 3, 4, and 5 were mixed and crossed, which implied that it was difficult to use SOM to estimate SFRB types 3, 4, and 5. One reason behind the phenomenon could be that the samples of these types are too few for the model to be trained adequately in the training stage. Another reason might be that SFRB of types 3, 4, and 5 have broad characteristics and belong to multiple types. Some obvious prediction errors were associated with SFRB where the original status has changed over time. For example, a disused drainage basin (type 3) may now be used as natural protection purposes (type 6).

From the perspective of feature reduction, it indicated that the 25 selected variables (see Table 5.4), which had high correlations with the SFRB types, were the important ones to characterize and distinguish SFRB, while the remaining ones were relatively redundant. Therefore, to some extent, this finding could lead to a reduction in sampling effort and cost.

Table 5.5: Prediction of SFRB types based on SOM analysis with 43/25 variables.

Real type	Predicted type	N-43/N-25	Real type	Predicted type	N-43/N-25
1 (4 sites)	1	0/0	3	5	3/1
2 (67 sites)	2	48/60	3	6	1/2
3 (14 sites)	3	4/1	4	3	0/1
4 (7 sites)	4	3/2	4	5	0/1
5 (19 sites)	5	7/7	4	6	4/3
6 (73 sites)	6	46/46	5	3	3/4
1	2	3/4	5	4	4/4
1	3	1/0	5	6	5/4
2	3	17/7	6	3	2/0
2	4	4/0	6	4	3/2
3	2	0/1	6	5	22/25
3	4	6/9			

Note: N-43 (The Number of the predicted SFRB based on 43 variables);
N-25 (The Number of the predicted SFRB based on 25 variables) .

To verify the effectiveness of the selected 25 variables, the SOM model was used to predict SFRB types by using all 43 characteristic variables. The prediction results were demonstrated in Table 5.5. For SFRB type 2, the prediction based on 25 variables achieved precision of 89.6% and the prediction based on 43 variables gained precision of 71.6%. However, their predictions of other types were very similar. For instance, it was still difficult to distinguish types of 3, 4 and 5 clearly. To summarize, the prediction of SFRB types based on 25 related variables outperformed that based on 43 variables. It might be because the more variables considered, the more environmental factors (i.e. higher variability) were involved and the more difficult it was to achieve high prediction accuracy. In general, the SOM model can perform well in predicting SFRB types 2 and 6 while poorly on predicting other types with the help of 25 highly related variables.

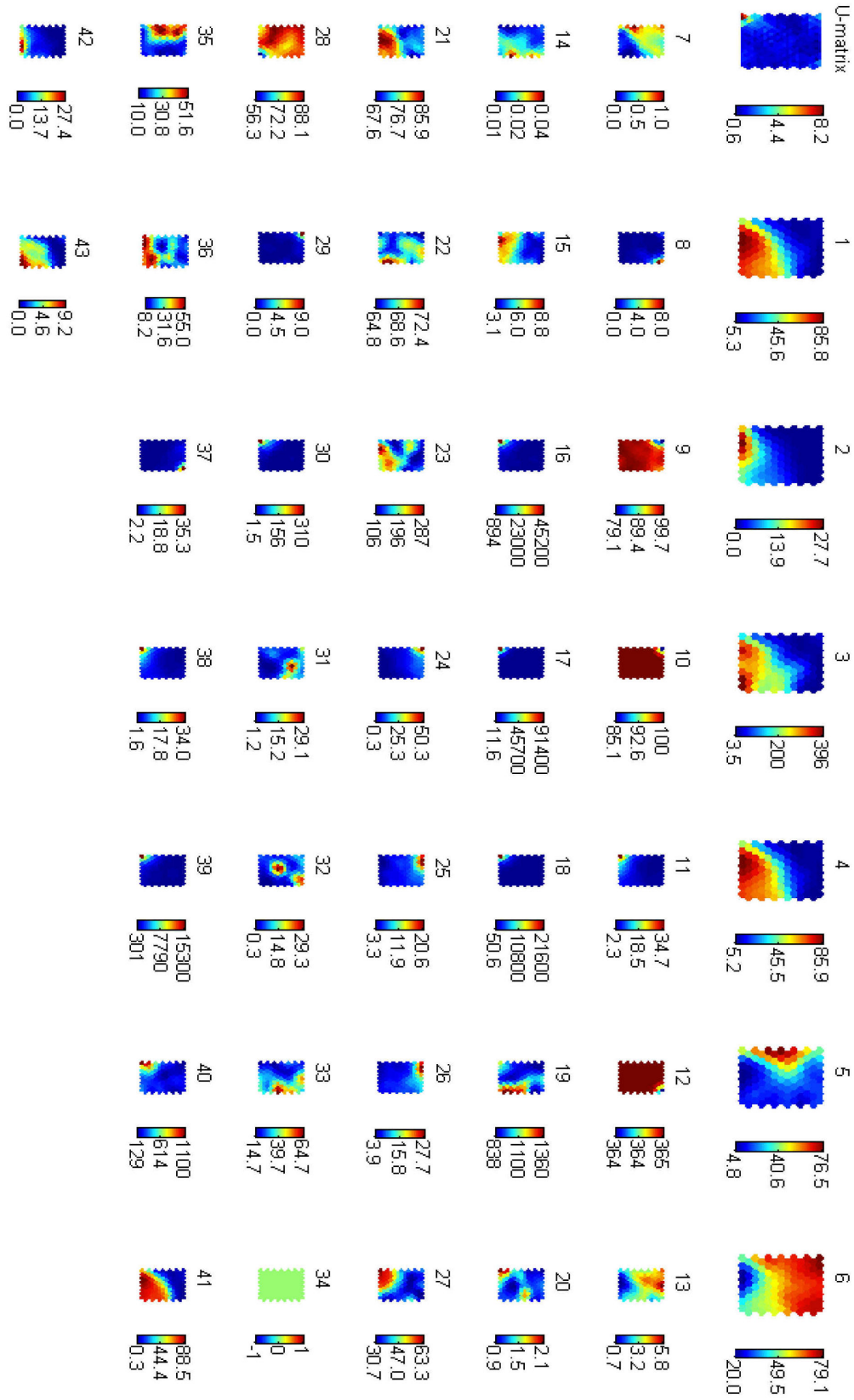
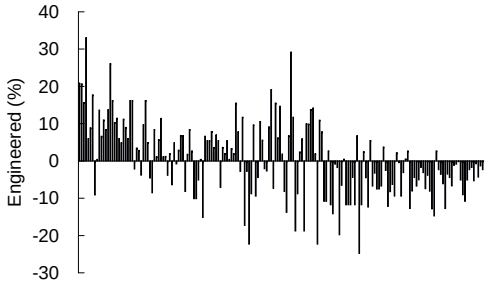
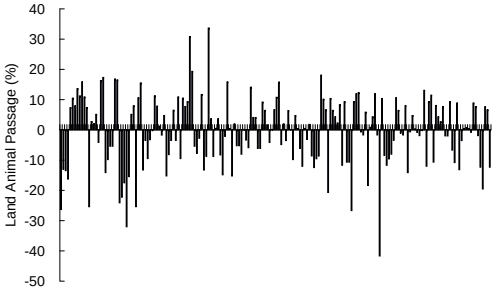


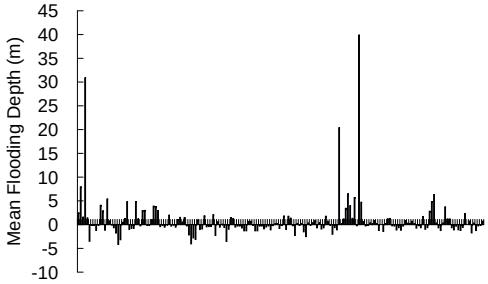
Figure 5.6: SOM Map of SFRB variables.



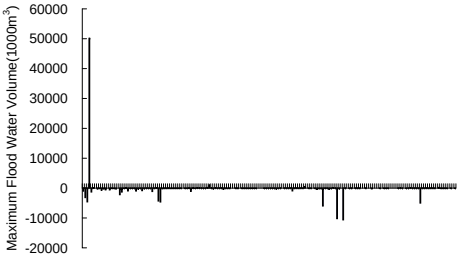
(a): Prediction of *Engineered*.



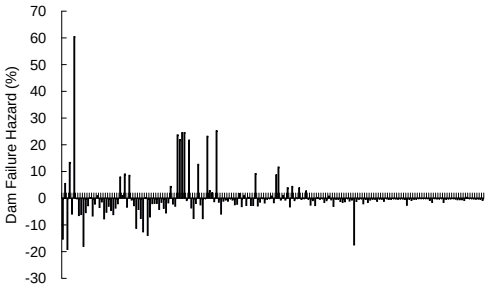
(b): Prediction of *Land Animal Passage*.



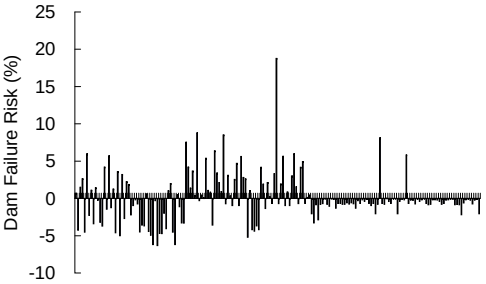
(c): Prediction of *Mean Flooding Depth*.



(d): Prediction of *Maximum Flood Water Volume*.



(e): Prediction of *Dam Failure Hazard*.



(f): Prediction of *Dam Failure Risk*.

Figure 5.7: The residual plot of the SFRB variables prediction based on SOM model.

5.4 Feature Selection on SFRB

5.4.1 SFRB Variables Selection

Three feature selection approaches Information Gain, Mutual Information, Relief (cf. Section 4.4.3) were used to order variables in terms of their importance to SFRB classes (types). However, the orders (sequences) of the variables were different for each feature selection method used. The findings of all three methods needed to be compared with each other to finally achieve a comprehensive list to be used for classification. For each variable, its sequence number in each method was obtained and then summed up. The lower the sum of sequence number gained, the more important the variable was. Consequently, the sum of sequence number was arranged in increasing order. The associated variables were ranked according to decreasing importance from top to bottom.

Table 5.6 provides an overview of the rankings generated by the three feature selection algorithms as well as the final rank. For example, taking variable 4 (*Outlet Arrangement and Operation*), its consequence numbers were 3, 2 and 2 for the methods Mutual Information, Relief and Information Gain, respectively. Thus, the sum of the consequence numbers was 7. Compared with the sum of consequence numbers for other variables, 7 was the second smallest number. While in terms of importance, variable 4 ranked second in the final order. The last column in Table 5.6 shows that variable 1 (*Engineered*) was the most important variable, variable 4 (*Outlet Arrangement and Operation*) was the second most important one and so on. Then, subsets of the ordered variables were passed to classifiers. This simple method, however, assumed that all feature selection methods were equally important and valid for SFRB.

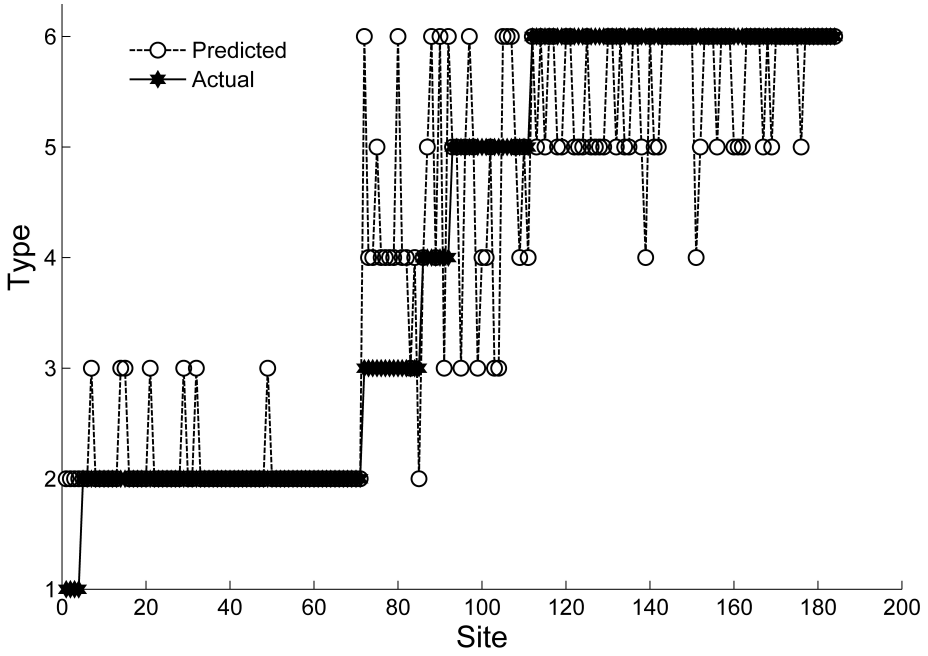


Figure 5.8: Prediction of sustainable flood retention basin types based on 25 variables

.

Table 5.6: Priority of selected variables based on three different feature selection algorithms. MI (Mutual Information), IG (Information Gain), SN (Sequence Number).

Variable	Sequence number of variables based on different algorithms			Sum of SN	Final Rank
	MI	Relief	IG		
1	1	1	1	3	1
2	7	3	3	13	3
3	6	4	4	14	4
4	3	2	2	7	2
5	11	5	7	23	7
6	8	6	6	20	6

Continued on next page

Table 5.6: (continued)

Variable	Sequence number of variables based on different algorithms			Sum of SN	Final Rank
	MI	Relief	IG		
7	5	7	5	17	5
8	35	9	8	52	16
9	27	13	9	49	14
10	37	15	11	63	20
11	21	11	10	42	10
12	34	12	17	63	21
13	15	18	16	49	15
14	22	21	19	62	19
15	13	16	15	44	11
16	38	22	18	78	26
17	36	23	12	71	23
18	39	24	20	83	28
19	16	10	22	48	12
20	4	14	14	32	9
21	9	8	13	30	8
22	29	20	25	74	24
23	10	17	21	48	13
24	12	19	23	54	17
25	17	25	24	66	22
26	26	26	26	78	27
27	32	27	29	88	33
28	19	28	30	77	25
29	28	33	36	97	18
30	30	30	27	87	31
31	2	29	28	95	35

Continued on next page

Table 5.6: (continued)

Variable	Sequence number of variables based on different algorithms			Sum of SN	Final Rank
	MI	Relief	IG		
32	14	31	39	84	29
33	31	32	37	100	37
34	33	38	35	106	39
35	20	34	33	87	32
36	18	35	32	85	30
37	25	36	34	95	36
38	24	37	31	92	34
39	40	39	38	117	40
40	23	40	40	103	38

The final rank of the variables showed the priority of the variables' importance on characterizing SFRB and distinguishing SFRB types. It means that the higher the rank of the variable is, the more contribution the variable makes to the SFRB classification, and thus the more important the variable is. In practice, the order of the key variables helps engineers, planners and practitioners to recognize which variables they should pay high attention during SFRB design, management and maintenance. For instance, compared with *Length of Basin* and *Width of Basin*, *Engineered* is of particular importance when engineers design a new SFRB site. If someone is designing a new SFRB of type 3 of any size, it is not necessary to construct high and long dams, which are usually associated with a SFRB of type 1. When planners decide to build a drinking water reservoir, they might consider *Engineered*, *Dam Length*, *Dam Height* and *Outlet Arrangement and Operation* as much more important than *Aquatic Animal Passage* and *Land Animal Passage*. For SFRB maintenance, practitioners should focus on how to maintain the outlet, dam structure and animal passage rather than how to maintain its groundwater infiltration or various catchment proportions. Acknowledgement of the importance of various variables reduces unnecessary costs and work effort.

5.4.2 SFRB Classification with Classifiers

To verify the effectiveness of the findings at the stage of feature selection (cf. Section 5.4.1), classification based on selected sub-sets of variables was performed. Four benchmark classifiers were applied during this process (cf. Section 4.4.4). Each time, the SFRB dataset was classified by one classifier, using different numbers of features (in order of decreasing priority) ranging from 1 to 40. Finally, the mean classification accuracy was compared between the four classifiers.

Figure 5.9 summarizes the classification performance of four classifiers based on the ranked features. Findings showed that the application of SVM, NaïveBayes and J48 classifiers (cf. Section 4.4.4) led to parallel patterns (from top to bottom: J48, NaïveBayes and SVM) and that the classification accuracy improved by increasing the number of variables from one to nine, but that the upwards trend reduced considerably afterwards until 35 variables were used. Although it was noticed that the 35th variable was important to improve the SFRB classification performance, it was associated with a lower ranking. This implied that the contribution of the 35th variable might be based on the combination of other variables that ranked before it (10th to 34th variables). This indicated that about 35 variables were sufficient for the classification, but that even nine variables had comparable classification accuracy. Taking SVM classifiers for instance, the classification accuracy rose from 77.6% to 86.7% by increasing the number of variables from one to nine, and then increased to the highest value of 89.4% when 35 variables were used. After that, the accuracy became slightly lower at about 89.2%.

The *K*NN classifier showed a different pattern from the above mentioned classifiers. The classification accuracy obtained the lowest classification value (81.4%) when only seven variables were used, but sharply achieved the highest classification accuracy of 87.1% when nine variables were applied. Afterwards, it dropped slightly with the increase of the number of variables.

It is obvious that only using the first nine variables could achieve sufficient classification accuracy, while introducing more variables did not necessarily improve the classification

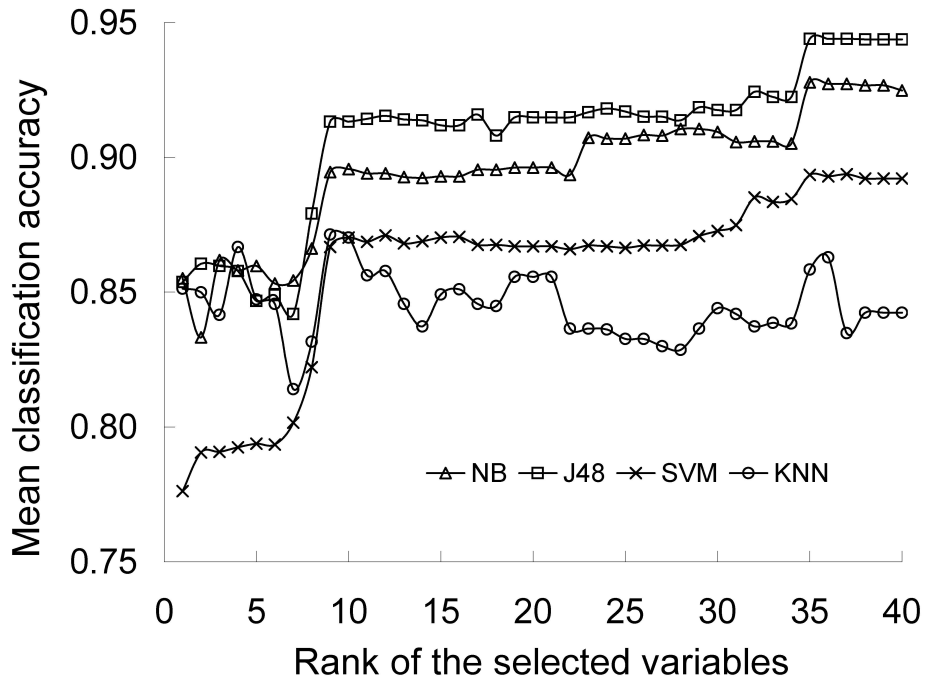


Figure 5.9: Comparison of the mean classification accuracies among the four classifiers. NB, NaïveBayes; J48; C4.5 Decision Tree; SVM, Support Vector Machine; KNN, K-Nearest Neighbours.

performance much. On the other hand, the accuracy could become worse. More specifically, Table 5.7 provides the classification accuracy of four different classifiers based on the first nine variables and the total forty variables. For all classification results, the 95% confidence interval has also been provided in Table 5.7. The classification results based on the first nine variables were 89.5% for NaïveBayes, 91.3% for J48, 86.7% for SVM and 87.1% for KNN. They have performed very well compared with the classification results of the total forty variables. The classification accuracy was only 3% lower for NaïveBayes, J48 and SVM. An improvement of 2.9% for KNN has been noted. Therefore, considering the contribution to SFRB classification, the first nine variables were regarded as the most important ones; they were as follows: *Engineered*, *Dam Height*, *Dam Length*, *Outlet Arrangement and Operation*, *Aquatic Animal Passage*, *Land Animal Passage*, *Floodplain Elevation*, *Impermeable Soil Proportion* and *Drainage*.

Table 5.7: Classification results for four classifiers based on the first nine variables and a total of forty variables (95% confidence intervals).

	NB	J48	SVM	*K*NN
First 9 variables				
Accuracy	89.5%	91.3%	86.7%	87.1%
Confidence interval	[85.76, 92.31]	[87.89, 93.92]	[82.78, 89.96]	[83.07, 90.20]
Total 40 variables				
Accuracy	92.5%	94.4%	89.2%	84.2%
Confidence interval	[89.12, 94.82]	[91.32, 96.36]	[85.46, 92.08]	[80.12, 87.79]

Note:NB (Naïve Bayes); J48 (C4.5 Decision Tree); SVM (Support Vector Machine); *K*NN (*K*-nearest-Neighbour).

5.4.3 Validation on Six Representative Case Studies

Theoretically, the selected nine important variables, which led to the high accuracy of SFRB classification, were supposed to distinguish SFRB types well. It followed that the most important variables were supposed to have very different characteristics for different SFRB types. For validation purposes, we explored the underlying relationships between the identified nine variables and the six SFRB types (see Table 5.8). One typical site for each SFRB type was studied to get insight into the behavior of the selected variables for individual types. Figure 5.10 to Figure 5.15 show pictures for the six representative cases, respectively.

Summary statistics showing the characteristics and relationships between different SFRB types, their functions and the identified variables were presented in Table 5.8. The results verified that the six types of SFRB could be distinguished successfully, even if only the identified nine key variables were used, as explained below for each case study. Findings clearly indicated, for example, that SFRB types 1 (predominantly used for hydraulic purpose) and 2 (mainly applied for drinking water supply) had highly engineered structures (98.6±0.9% and 69.8±10.1%, respectively) with high dams and advanced outlet arrangements such as spillways (97.7±1.0% and 69.5±11.0%,

Table 5.8: Summary statistics (mean $\mp$ standard deviation) characterizing the relationships between SFRB types, functions and the nine key variables.

KV	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
E	98.6±0.9	69.8±10.1	26.9±8.0	25.4±8.8	25.7±9.9	5.7±3.9
DH	25.8±19.1	11.0±8.4	1.8±1.2	2.0±0.9	1.1±1.3	0.0±0.2
DL	277.7±172.4	277.0±238.1	113.6±100.5	75.0±59.4	97.9±206.9	0.9±6.2
OAO	97.7±1.0	69.5±11.0	25.8±10.9	25.0±9.6	19.2±14.5	5.4±4.6
AAP	0.0±0.0	13.1±13.9	30.8±30.1	30.0±16.6	26.4±28.1	38.2±35.8
LAP	11.7±3.5	45.4±16.9	66.8±9.7	65.4±9.9	67.4±10.9	71.0±10.3
FE	0.0±0.0	0.2±0.4	0.6±0.4	0.6±0.3	0.6±0.3	0.7±0.4
D	1.2±0.2	1.1±0.5	2.2±0.4	1.2±0.2	1.0±0.4	1.1±0.5
ISP	81.3±5.5	82.8±7.5	64.5±7.7	74.2±9.8	75.1±7.4	71.7±8.9
PF	Hydraulic purpose	Drinking water supply	Sustainable drainage	Aesthetic landscape and sustainable drainage	Recreation	Environment protection
CSE	Lubreoch power station	Harperrig Reservoir	DEX wetland	Caw burn wetland	Lanark Loch	Hare Myre
GRCS	$56°32'N$; $4°35'W$	$55°50'N$; $3°27'W$	$56°04'N$; $3°24'W$	$55°55'N$; $3°29'W$	$55°40'N$; $3°45'W$	$56°34'N$; $3°20'W$

Note: KV (Key Variables); E (*Engineered*, %); DH (*Dam Height*, m); DL (*Dam Length*, m); OAO (*Outlet Arrangement and Operation*, %); AAP (*Aquatic Animal Passage*, %); LAP (*Land Animal Passage*, %); FE (*Floodplain Elevation*, m); D (*Drainage*, cm/d); ISP (*Impermeable Soil Proportion*, %);PF (Predominant Functions); CSE (Case Study Examples); GRCS (Grid Reference for Case Study); ADEX (Dunfermline Eastern Expansion).

respectively), while type 6 was rather natural (5.7±3.9% for *Engineered*) without a dam and designed outlets. Concerning SFRB types 1 and 2, the man-made and therefore highly engineered structures resulted in barriers for animals. Therefore, both *Aquatic Animal Passage* and *Land Animal Passage* were low. While for SFRB type 6 (mainly for the purpose of environmental protection) characterized by natural properties, the animal movement was not severely restricted (38.2±35.8% and 71.0±10.3% for *Aquatic Animal Passage* and *Land Animal Passage*, respectively). SFRB types 3 (mainly used for sustainable drainage), 4 (predominantly applied for landscape enhancement) and 5 (mainly used recreational activities) had similar characteristics such as low engineered structures; for example, low dams (about 3 to 5 m high) and poor outlet arrangements (potentially only weirs present). However, SFRB type 3 was distinctive from the

other types due to its high values for *Drainage* (2.2±0.4 cm/d). Correspondingly, the *Impermeable Soil Proportion* for type 3 was relative low. SFRB type 4 had slightly higher *Drainage* values than type 5. Furthermore, water bodies that are mainly used for wastewater treatment belong to type 4 while those used in parks or for other recreational activities are regarded as type 5. More specific characteristics associated with the nine key variables for the six representative case studies are discussed below.

The SFRB of type 1 are mainly hydro-electric power stations and current drinking water reservoirs. Generally, they are precisely designed, well maintained and automatically controlled. So, this kind of SFRB tends to obtain relatively high values for *Engineered*, *Dam Height*, *Dam Length*, *Outlet Arrangement and Operation* and *Impermeable Soil Proportion*, but very low scores for the variables Aquatic Animal Passage, Land Animal Passage, Drainage and Floodplain Elevation. For example, Lubreoch power station (Figure 5.10) is one typical SFRB type 1 case study. Being built for power generation, it got a high score for *Engineered* (98%). Meanwhile, the *Dam Height* (39 m), *Dam Length* (530 m) and *Outlet Arrangement and Operation* (97%) were also very high. Due to its steep concrete spillway and high dam structure, the passage of aquatic and land animals are severely hampered by the physical structures of this SFRB. For this reason, both *Aquatic Animal Passage* and *Land Animal Passage* obtained relatively low scores (close to 0). Since the power station is fully controlled and a large spillway releases extra water via an overflow, there is no clearly defined *Floodplain Elevation*. Water does not normally penetrate the dam or drains easily from the basin. So the *Impermeable Soil Proportion* for this site was relatively high at around 90% and the *Drainage* was fairly low at 1.3 cm/d.

Comparatively, the dam and spillway tends to be of medium size for SFRB type 2. They generally have high values for *Engineered* (60% to 85%), *Impermeable Soil Proportion* (70% to 90%) mean values for *Dam Height*, *Dam Length* and *Outlet Arrangement and Operation* (50% to 75%), but low values for the *Aquatic Animal Passage*, *Land Animal Passage* and *Floodplain Elevation*. For example, as a drinking water supply reservoir with a dam (14 m high and 154 m long) and a spillway, Harperrig Reservoir (Figure 5.11) was assigned 85% for *Engineered* and 75% for *Outlet Arrangement and Operation*.

Figure 5.10: Loch Lyon is a typical example of a Hydraulic Flood Retention Basin (Sustainable Flood Retention Basin type 1).

For the purpose of keeping water within the drinking water reservoir, the *Impermeable Soil Proportion* is as high as 90% and *Drainage* is as low as 0.5 cm/d. Due to the presence of a high dam structure and the lack of a fish ladder, the *Aquatic Animal Passage* and *Land Animal Passage* values are 10% and 50%, respectively. A spillway exists and has just been modified to optimize the flood protection purpose.

The mature Dunfermline Eastern Expansion (DEX) wetland (SFRB type 3; Figure 5.12)) characterized by dense stands of reeds was assigned a low value for the variable *Engineered* (35%) since it is influenced by the backwater which has a active control structure nearby (not shown on the picture). The outlet is a small weir with a width of 2.5 m. Several large stones are located in front of the weir to dissipate energy. Consequently, *Outlet Arrangement and Operation* and *Aquatic Animal Passage* were assigned 20% and 60%, respectively. Considering that the wetland system is located within a public park, this may result in less of a barrier for land animals. It follows that the wetland obtained 80% for *Land Animal Passage*. According to the topography

Figure 5.11: Harperrig Reservoir is a typical example of a Traditional Flood Retention Basin (Sustainable Flood Retention Basin type 2).

of the DEX area, the *Floodplain Elevation* was estimated to be roughly 1 m. The *Impermeable Soil Proportion* was estimated to be 85%, since the area is predominantly characterized by clay soil of low permeability. Considering the high water retention potential within the wetland, the *Drainage* rate was estimated to be 2 cm/d.

The Caw Burn wetland (SFRB type 4; Figure 5.13) is operated as an off-line structure treating the base flow and first foul flush of an urban area only. The wetland is fed by an abstraction pipe and overflows into a swale leading to the receiving watercourse. No engineered structures were constructed at the outlet of the wetland. A relatively low value of 20% was therefore assigned to the variable *Engineered* and 5% was assigned

Figure 5.12: Dunfermline Eastern Expansion is a typical example of a Sustainable Flood Retention Wetland (Sustainable Flood Retention Basin type 3).

to *Outlet Arrangement and Operation*. The Caw Burn wetland is separated from the receiving watercourse by a shallow earth dam of 0.5 m height and 150 m width. However, the earth dam overflows at several locations due to its shallow nature during storm events. Since there are no engineered barriers for animals, this SFRB obtains a high score for both *Aquatic Animal Passage* (80%) and *Land Animal Passage* (80%). The *Floodplain Elevation* is around 0.1 m during the dry season. Considering the surrounding soil, which is dominated by old shale mining waste, the *Impermeable Soil Proportion* is estimated to be 70%. The Caw Burn keeps the wetland permanently wet. The *Drainage* rate was estimated to be as low as 0.1 cm/d.

The value of *Engineered* assigned to Lanark Loch (SFRB type 5; Figure 5.14) was 25%. The dam height and length were 1 m and 20 m, respectively. The *Outlet Arrangement and Operation* was also low (15%) since only a small weir and a pipe exist near the outflow. Due to the low control of the outflow, the *Aquatic Animal Passage* was relatively high (40%). The variable *Land Animal Passage* also obtained

Figure 5.13: Cawburn Wetland is a typical example of an Aesthetic Flood Treatment Wetland (Sustainable Flood Retention Basin type 4).

a relatively high value (70%), because there are no obvious obstacles preventing land animals from roaming. Due to the flat topography of the basin area and the small capacity of the basin, the value for the *Floodplain Elevation* was also low. The estimated value of *Impermeable Soil Proportion* was 70% and *Drainage* was 0.8 cm/d (based on an assessment of the soil present at the site).

Water bodies belonging to SFRB type 6 are relative natural without high engineered structures such as dams, spillways or sluice gates. They usually obtain very low values for variables such as *Engineered*, *Outlet Arrangement and Operation* and *Impermeable Soil Proportion*. Taking the Site of Specific Scientific Interest Hare Myre (Figure 5.15), for example, 2% was given to the variable *Engineered*. Accordingly, *Outlet Arrangement and Operation* also received a score as low as 3%. There is no dam present, so the variables *Dam Length* and *Dam Height* are not applicable but values of zero each were noted. Without the presence of an outflow, *Aquatic Animal Passage* was therefore also zero. Due to lack of obstacles around the wetland, *Land Animal Passage* is given a

Figure 5.14: Lanark Loch is a typical example of an Integrated Flood Retention Wetland (Sustainable Flood Retention Basin type 5).

high value of 80%. In contrast with drinking water reservoirs (usually SFRB types 1 or 2), there are no man-made structures to protect water from draining and infiltrating. Based on the observation of the soil of the basin, 50% and 1.5 cm/d were assigned to *Impermeable Soil Proportion* and *Drainage*, respectively. According to the topography of the catchment of the basin, the *Flood Elevation* was estimated at 0.5 m.

The effectiveness of the nine key variables selected according to feature selection in Section 5.3.1 was successfully verified on six representative SFRB case studies. Reducing the number of the surveyed variables from forty to nine saves time, money and labor resources for the assessment of SFRB and achieves a rapid classification of SFRB. Furthermore, due to different types of SFRB having different performances regarding the nine classification variables, the rapid classification is beneficial in providing engineers, practitioners and land use planners with scientific support during design, maintenance and decision making. For example, the values of *Engineered*, *Dam Height*, *Dam Length* and *Outlet Arrangement and Operation* should be high when someone is

Figure 5.15: Hare Myre is a typical example of a Natural Flood Retention Wetland (Sustainable Flood Retention Basin type 6).

designing a SFRB of type 1, but they should be low for SFRB of type 6. Regarding the latter, which should have low values for the variable *Engineered*, practitioners could consider how to improve the biological integrity of the SFRB rather than how to increase its water supply capability. For SFRB of type 4, authorities should focus on how to improve water quality but not flood reduction capability. If the variable *Engineered* for a very old drinking water reservoir decreases over time and its outlet does not work properly any more, planners might consider to assign a new SFRB status such as SFRB of type 5 used for public recreation. Furthermore, some drinking water reservoirs can also be used for flood control purposes by actively controlling their water levels. Therefore, with the aid of the linkage between the result of SFRB classification and the identified functions of SFRB, engineers' design of new SFRB may better meet practical needs, authorities might gain an improved understanding of maintenance requirements, and planners may consider more comprehensively how to manage the development of future SFRB.

5.5 Multi-label Classification of SFRB

5.5.1 Classification Results

The multi-label algorithms MLSVM, MLKNN and BP-MLL (cf. Section 4.5.3) were performed on two data sets comprising 372 and 202 SFRB (associated with 43 variables each) located in Scotland and Baden, respectively. For each method, one-leave-out cross-validation was used in the training phase and five loops were tried for each individual parameter. This procedure makes the parameters less sensitive to classification results. Moreover, the performances of the multi-label classifiers were evaluated by using the five metrics in Section 4.5.4.

Table 5.9 summarizes the classification results of each classifier for both data sets. For the Scottish data set, all the three multi-label classifiers achieved good results. Specifically, the average precisions of the MLSVM, MLKNN and BP-MLL techniques were 91.8 ± 1.1%, 92.1 ± 3.3% and 90.1 ± 19.8%, respectively. Other criteria such as *One-error*, *Ranking-Loss* and *Hamming-Loss* were all small enough, which confirms the good performance. The corresponding values for the German data were slightly higher; i.e. 91.9 ± 1.0%, 95.9 ± 1.3% and 94.2 ± 11.9% for MLSVM, MLKNN and BP-MLL, respectively. Furthermore, the *Hamming Loss* and *One-error* values for the data set from Baden were all smaller than those for the Scottish data set. Although multi-label learning is more complicated, the good classification results indicated that all three multi-label classifiers can predict the types of SFRB effectively.

Furthermore, for comparison, Table 5.9 illustrates the classification results for the three traditional learning algorithms SVM, KNN and BP applied on both data sets in terms of *Accuracy*. However, since multi-label classification and traditional classification use different evaluation metrics, as stated in Section 4.5.4, it is difficult to compare the classification results directly with each other. The reason is that the accuracy is defined to evaluate traditional (single label) classification in which the predicted class is either correct or not, while *Average Precision* is used for evaluation of multi-label classifiers in

Table 5.9: Experimental results based on multi-label learning algorithms and traditional learning algorithms (mean ± standard deviation).

Data set	Algorithm for SFRB in Scotland (372 sites)		
Evaluation Criteria	MLSVM	MLKNN	BP-MLL
Average Precision	0.918±0.011	0.921±0.033	0.901±0.198
Coverage	0.878±0.050	0.833±0.149	0.839±1.039
One-error	0.105±0.018	0.102±0.044	0.151±0.359
Ranking-Loss	0.066±0.009	0.062±0.025	0.069±0.150
Hamming-Loss	0.111±0.017	0.109±0.036	0.163±0.151
	SVM	*K*NN	BP
Accuracy	0.861±0.020	0.830±0.006	0.883±0.002

(Continue)

Data set	Algorithm for SFRB in Baden (202 sites)		
Evaluation Criteria	MLSVM	MLKNN	BP-MLL
Average Precision	0.919±0.010	0.959±0.013	0.942±0.119
Coverage	1.083±0.063	0.721±0.108	0.856±1.232
One-error	0.014±0.005	0.020±0.000	0.020±0.140
Ranking-Loss	0.084±0.010	0.035±0.014	0.054±0.109
Hamming-Loss	0.077±0.009	0.082±0.021	0.083±0.114
	SVM	*K*NN	BP
Accuracy	0.900±0.002	0.905±0.000	0.897±0.010

Note: MLSVM (Multi-Label Support Vector System); MLKNN (Multi-Label K-Nearest Neighbour); BP-MLL (Back-Propagation for Multi-Label Learning); SVM (Support Vector System); KNN (K-Nearest Neighbour); BP (Back-Propagation).

which the predicted labels are either fully correct, partly correct or fully incorrect. Only if all types of a SFRB are predicted, the classification is regarded as fully correct, which

is thus stricter than accuracy to some extent. Although the *Accuracy* and *Average Precision* have different definitions, it is interesting to note that the performances of traditional classifications were weaker than those of the multi-label methods based on the experiments. Specifically, three traditional classifiers: SVM, KNN and BP obtained accuracies of 86.1 ± 2.0%, 83.0 ± 0.6% and 88.3 ± 0.2% respectively on Scottish data. In comparison, the *Average Precision* values of the three multi-label classifiers were all above 90%. For German SFRB data, the classifiers SVM, KNN and BP achieved an accuracy of 90.0 ± 0.2%, 90.5 ± 0.0% and 89.7 ± 1.0% respectively compared with the *Average Precision* of multi-label classifiers ranging from 91.9% to 95.9%. From the experimental results, it is obvious that multi-label classification algorithms yield better results. The benefits of the multi-label classification models over traditional methods will be further illustrated by three representative case studies in the following Section 5.5.2.

The comparative findings show that multi-label classification provides a more robust and efficient tool to better classify and understand SFRB than the traditional methods. This tool allows one SFRB to belong to several types simultaneously. Multiple types associated with a single SFRB verify what is observed in practice; i.e. one SFRB often performs multiple functions. Thus, multi-label classification better represents the complex reality than traditional approaches. Furthermore, this tool provides new insights to help people understand the true status and multiple functions of SFRB in a comprehensive way, avoiding conflicts and confusions about SFRB assessment between engineers, stakeholders and planners. For example, for a natural water body partially developed for water sports, some people may state it belongs to SFRB type 5 due to the recreations, while others insist it belongs to SFRB type 6 because of its natural characters. Conflicts may be avoided when it is judged by planners and decision-makers as both type 5 and 6 simultaneously by multi-label classifiers, as the public can recognize that the SFRB has multiple and justified functions at the same time.

The findings of multi-label classification also provide valuable support on decision making for planners, designers and practitioners.When making plans for the development and management of SFRB, realizing the multiple functions of SFRB, planners might

consider the relationships and cooperation of the functions rather than focusing only on the purpose of SFRB that was originally recognized. Similarly, some designers may promote more people-friendly rather than highly engineered facilities. For a SFRB originally used for sustainable drainage (SFRB type 3) but also with an aesthetic landscape, practitioners will not ignore its aesthetics when considering maintenance. Moreover, multi-label classifiers may estimate desirable functions of SFRB which are currently not noticed, therefore, the planners and practitioners can develop the promising functions properly in the future and generate new potential benefits. For example, a large number of drinking water reservoirs (SFRB type 2) located in central Scotland are currently used (and only used) for water supply purposes. Multi-label classifiers estimated them as type 1 as well which are used for flood control. Thus, the planners and practitioners can think about the existing reservoirs' potential use as flood defence structures, which will benefit society and save much money. In brief, multi-label classification can guide people, find and understand new functions of SFRB besides its initial purposes and then make better decisions to develop and manage SFRB.

5.5.2 Representative Case Studies

The relationship between multi-label classification and SFRB functionality will be demonstrated with the help of three relevant examples in this section. Furthermore, this section also shows the need for multi-label classification and its application value in practice. The multi-label classification of SFRB performed very well under the scheme of multi-label learning. It allowed one SFRB to belong to multiple types simultaneously, which helped people consider the development of SFRB from a multi-disciplinary and holistic perspective.

Johnston Loch ($55.89°N$, $4.09°W$) shown in Figure 5.16 is located in North Lanarkshire, Scotland. It drains the runoff from the surrounding residential area called Gartcosh. Additionally, it is an ideal place for walking and relaxing because of its peaceful and beautiful natural sights. The overall structure of the loch is entirely natural; i.e. there are currently no engineered structures like a dam or a spillway. Thus, it can be

intuitively characterized as a SFRB of type 3 (mainly used for sustainable drainage), 5 (largely used for aesthetic recreation), and 6 (predominantly includes natural water bodies). Johnston Loch was predicted as SFRB types 3, 5 and 6 by both classifiers MLKNN and BP-MLL, while MLSVM predicted it as SFRB types 5 and 6. However, Johnston Loch was judged as SFRB type 6 only by the traditional classifiers SVM, KNN as well as BP. Therefore, the predictions of the multi-label classifiers represented the diverse functions of Johnston Loch well.

The results of the traditional classifications were too limited. The findings may mislead people, which could lead to a neglect of the loch's drainage and aesthetic functions. Realizing the fact that Johnston Loch has multiple functions, planners and designers can develop the basin more sustainably and cost-effectively.

Figure 5.16: Johnston Loch ($55.89°N$, $4.09°W$) is located in Gartcosh, Scotland.

Taking Murg Ausgleichsbecken ($48.40°N$, $8.21°W$; Figure 5.17) located in Forbach (Baden, Germany) as another random example. It is a purpose-built SFRB used for flood protection. However, the integrative functions of enhancing the landscape aesthetics and recreational activities have been growing in importance after its construction.

Therefore, Murg Ausgleichsbecken was characterized not only as a SFRB type 1 but also as a SFRB type 5. All three classifiers MLSVM, MLKNN and BP-MLL predicted the basin as types 1 and 5 simultaneously. This confirms the findings of experts who visited this SFRB in 2006 and 2010. Realizing that the Murg Ausgleichsbecken actually belongs also to type 5 using the new classification method, planners might enhance its aesthetic and recreational attributes by providing wider public access. Without the help of multi-label classification, planners might only focus on the flood control purpose, and thus loose many valuable societal benefits of the basin.

Figure 5.17: Murg Ausgleichsbecken ($48.40°N$, $8.21°W$) is located in Forbach (Baden, Germany).

The previously introduced Harlaw Reservoir (Figure 4.3, Section 4.5.1) was also correctly predicted as SFRB type 1(mainly used for flood control) and type 2 (predominantly used for drinking water supply) by the multi-label classifiers. The prediction results verify the fact that the reservoir plays an important role in flood defence by adjusting runoff release quantities. More importantly, realizing that reservoirs and other basins can be adapted to contribute to flood control will save huge sums amounts of

money and resources that can be spend elsewhere on new flood defence structures. If somebody can clearly identify an SFRB as being purely of type 1 and if it is subsequently managed properly, the structure is likely to contribute substantially to sustainable flood risk management planning.

These representative case studies indicate that the results of multi-label classification are comprehensive and much closer to the real situations of the basins. To some extent, traditional classification is limited to assessing each SFRB from a single perspective, while multi-label classification has broader horizon on assessing SFRB. Multi-label classification has obvious advantages over traditional classification in providing planners, designers and practitioners with efficient and reliable information with respect to SFRB management and development. Helping people understand SFRB from multiple perspectives, identification of multiple functions can avoid the conflicts and confusions of SFRB development among planners, designers and practitioners. Proper management of each SFRB, based on multi-label classification, will make most use of its functions and optimize its benefits to society.

5.6 Spatial Analysis

5.6.1 Statistics of Flood Related Variables

As stated in Section 4.6.4, the spatial research only focused on the flood-related variables such as *Engineered*, *Mean Flooding Depth*, *Managed Mean Flooding Depth*, *Maximum Flood Water Volume* and *Managed Maximum Flood Water Volume*. Before the spatial analysis of the SFRB data set, these flood-related variables were analyzed. Table 5.10 shows the summary statistics of the original values of these variables. The data variability for most variables was relatively high, reflecting the diversity of SFRB. The high variability of *Managed Maximum Flood Water Volume* is likely to negatively influence the prediction errors of kriged maps (see Section 5.6.2 and Section 5.6.3).

Table 5.10: Summary statistics of the flood-related variables.

Statistic	E	MFD	MMFD	MFWV	MMFWV
Minimum	0	1	0	2450	0
Maximum	100	70	42	260000×10^4	14200×10^4
Mean	40.5	6.9	4.5	2312×10^4	462×10^4
Standard deviation	33.52	8.13	6.01	18801×10^4	1589×10^4

Note: E (*Engineered*, %); MFD (*Mean Flooding Depth*, m); MMFD (*Managed Mean Flooding Depth*, m); MFWV (*Maximum Flood Water Volume*, m^3); MMFWV (*Managed Maximum Flood Water Volume*, m^3).

5.6.2 Findings based on Ordinary Kriging

Ordinary kriging (cf. Section 4.6.3) was applied for the key flood control variables *Engineered*, *Mean Flooding Depth*, *Managed Mean Flooding Depth*, *Maximum Flood Water Volume* and *Managed Maximum Flood Water Volume*. The statistics of these target variables are summarized in Table 5.10. The high variance of these variables reflects the diverse types of SFRB. Table 5.11 presents the ordinary kriging characteristics for these variables.

Figure 5.18 to Figure 5.22 show examples of applying ordinary kriging for the variables *Engineered*, *Mean Flooding Depth*, *Managed Mean Flooding Depth*, *Maximum Flood Water Volume* and *Managed Maximum Flood Water Volume* respectively. High numerical values for the variable *Engineered* generally indicate the likely necessity for high civil engineering investment to be made when planning for the construction of a new SFRB (Figure 5.18). The most engineered SFRB structures are likely to be found in the south-west of the study area, which coincides with the highest density of reservoirs and lakes used for water supply purposes. In contrast, for the study area in the north, relatively low investment for flood infrastructure is required. This variable is particularly useful when a decision has to be made on where old flood infrastructure needs to be upgraded or new SFRB constructed. It would be preferable to have areas

Table 5.11: Summary of ordinary kriging characteristics for the flood related variables.

Variables		Variogram parameter			Variogram error
	Model	Range	Partial sill	Nugget	MSE of prediction
E	Exponential	125534	255.04	953.06	−0.0018
MFD	Gaussian	69788	52.19	25.26	−0.0103
MMFD	Spherical	96548	24.3	17.59	−0.0036
MFWV	Spherical	167522	3.6297×10^{16}	1.5532×10^{16}	0.0044
MMFWV	Gaussian	167522	2.5236×10^{14}	1.5356×10^{4}	0.0072

Note: MSE (Mean Standardized Error); E (*Engineered*, %); MFD (*Mean Flooding Depth*, m); MMFD (*Managed Mean Flooding Depth*, m); MFWV (*Maximum Flood Water Volume*, m^3); MMFWV (*Managed Maximum Flood Water Volume*, m^3).

prone to flooding being located within catchments associated with low values for the variable *Engineered* and high values for those such as *Managed Mean Flooding Depth* and *Managed Maximum Flood Water Volume* (see below).

The spatial distribution for the variables *Mean Flooding Depth* and *Managed Mean Flooding Depth* are shown in Figure 5.19 and Figure 5.20 respectively. The *Mean Flooding Depth* is relatively high in the less populated upland areas of the north-west and south of the study area as well as within the Pentland Hills, a small area directly located south-west of Edinburgh (Figure 2.7). Low values for the variable *Mean Flooding Depth* are rare and patchy.

In comparison, the *Managed Mean Flooding Depth* variable is high only in the north-east and south-west of the study area. Moreover, high values have also been noted for some parts of the Pentland Hills. The area situated to the south-west of Edinburgh also has the highest density of reservoirs that could be used for hydraulic purposes such as flood control to protect the capital. The south-east and north-central regions are dominated by low flooding depths. The comparison indicates that the new variable

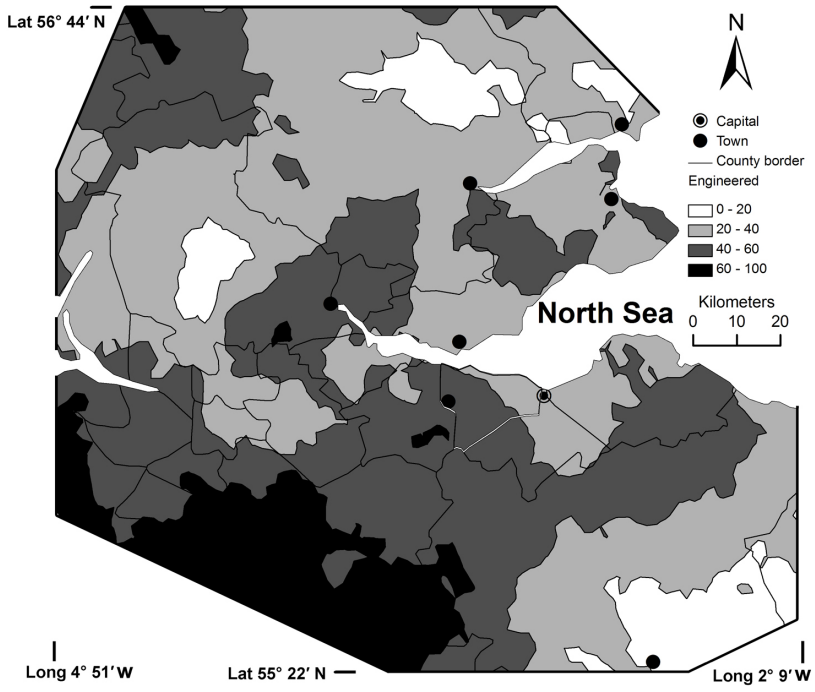


Figure 5.18: Ordinary kriging for *Engineered* (%).

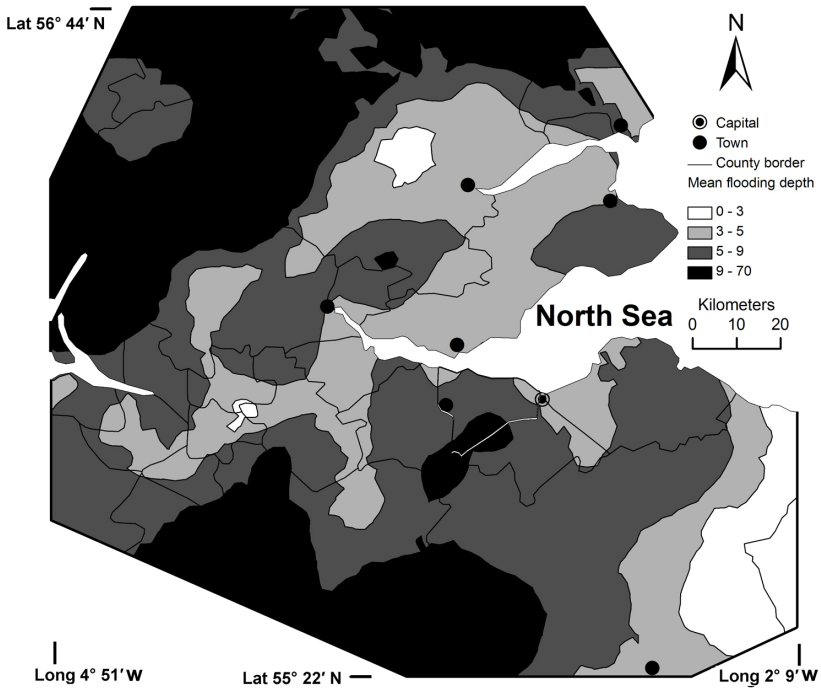


Figure 5.19: Ordinary kriging for *Mean Flooding Depth* (m).

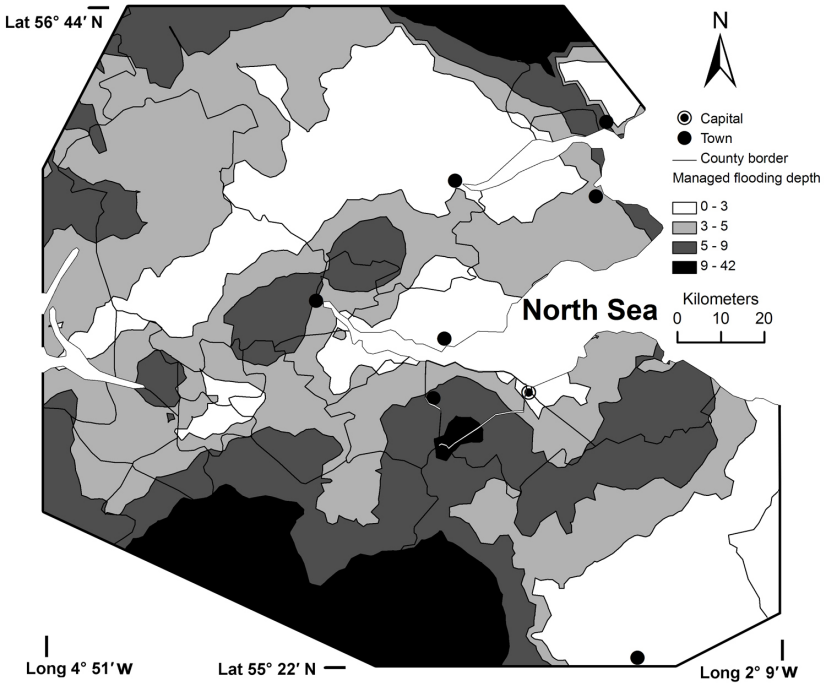


Figure 5.20: Ordinary kriging for *Managed Mean Flooding Depth* (m).

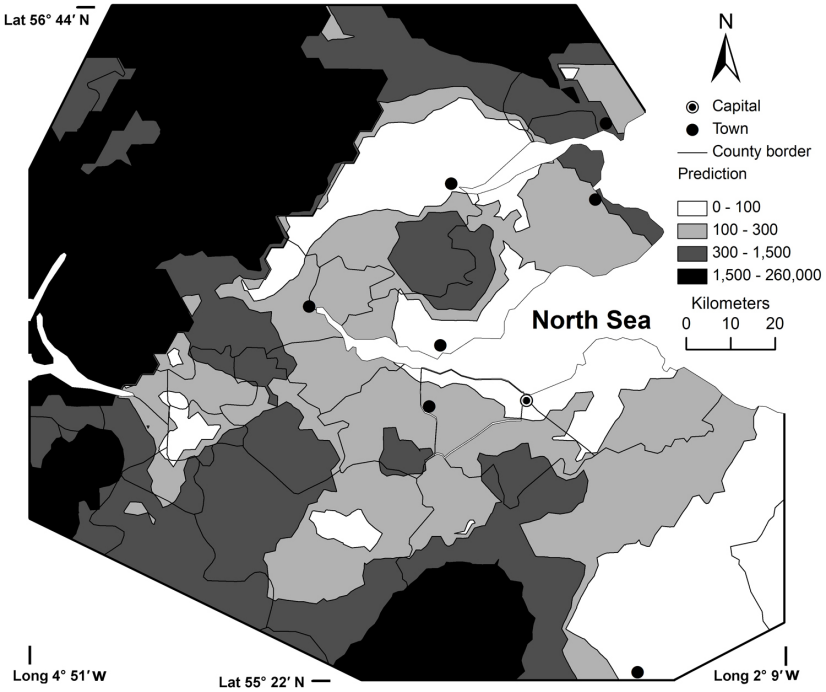


Figure 5.21: Ordinary kriging for *Maximum Flood Water Volume* (m^3).

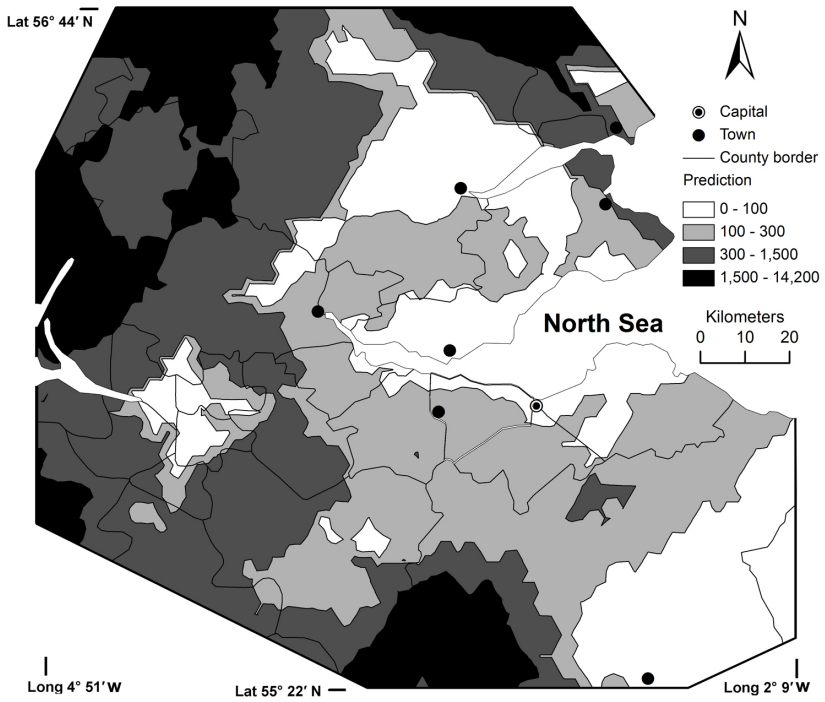


Figure 5.22: Ordinary kriging for *Managed Maximum Flood Water Volume* (m^3).

Managed Mean Flooding Depth has much lower values than the old variable *Mean Flooding Depth*, because the permanent water contained in lakes has not been taken into account when calculating the former variable. The new variable is therefore a much better indicator for high flood control potential.

Figure 5.21 and Figure 5.22 show the most likely values for the variables *Maximum Flood Water Volume* and *Managed Maximum Flood Water Volume*. These volume-based variables mirror the depth-based variables, indicating that higher depths relate to higher volumes, which is particularly the case for upland areas far away from major urban settlements.

Ordinary kriging has proved to be useful for more than 90% of the SFRB ground surveys, because it results in maps which are easy to understand for practitioners (e.g. Figure 5.18 to Figure 5.22). Maps appear to be smooth because most short-range noise was removed during kriging, which allows the flood risk manager to identify the key underlying patterns within complex data structures.

5.6.3 Findings based on Disjunctive Kriging

The maps produced with ordinary kriging are helpful in identifying areas of low or high values for various SFRB variables. The findings can be used directly by flood risk managers and landscape planners. However, these maps should not be taken at face value since many are misleading to a greater or lesser extent. Therefore, for SFRB designers and maintainers, it is necessary to use a technique such as disjunctive kriging (cf. Section 4.6.3), which provides estimates of the probability based on data given that the true values exceed a threshold at an unsampled location. It is need to perform a lognormal transformation for some variables in order not to violate underlying assumptions for disjunctive kriging.

Table 5.12: Summary of disjunctive kriging characteristics for the flood related variables.

Variables	Variogram parameter						Error
	Transformation	Model	Range	Partial sill	Nugget	Pt	MSE
E	normal score	exponential	16663	0.32	0.61	30	0.0173
MFD	none	exponential	34300	15.16	50.83	3	0.0422
MMFD	normal score	spherical	15806	0.41	0.48	3	0.0351
MFWV	log	spherical	61058	0.79	4.36	3.5×10^5	0.0988
MMFWV	normal score	spherical	37555	0.22	0.74	3.5×10^5	-0.0098

Note: MSE (Mean Standardized Error); Pt (Primary threshold); E (*Engineered*, %); MFD (*Mean Flooding Depth*, m); MMFD (*Managed Mean Flooding Depth*, m); MFWV (*Maximum Flood Water Volume*, m^3); MMFWV (*Managed Maximum Flood Water Volume*, m^3).

Table 5.12 shows a summary of disjunctive kriging parameters for the key flood control variables *Engineered*, *Mean Flooding Depth*, *Managed Mean Flooding Depth*, *Maximum Flood Water Volume* and *Managed Maximum Flood Water Volume*. Based on data properties and expert judgment derived from internal research team based on literatures, the thresholds for the variable Engineered and any variables indicating

flooding depth and flood water volume were set as 30%, 3 m and 350000 m^3 respectively. Similar to ordinary kriging, the model that led to minimum mean standardized error was selected as the most suitable model fitting the variables.

The examples showing the application of disjunctive kriging for the key variables are summarized in Figure 5.23 to Figure 5.27. Areas of low and high probabilities for the variable Engineered are relatively small and patchy (Figure 5.23). The probability map shown in Figure 5.23 can be used in conjunction with Figure 5.18 and all maps indicating flooding depth and flood water volume to determine the areas of greatest investment potential if flooding is likely to be a problem.

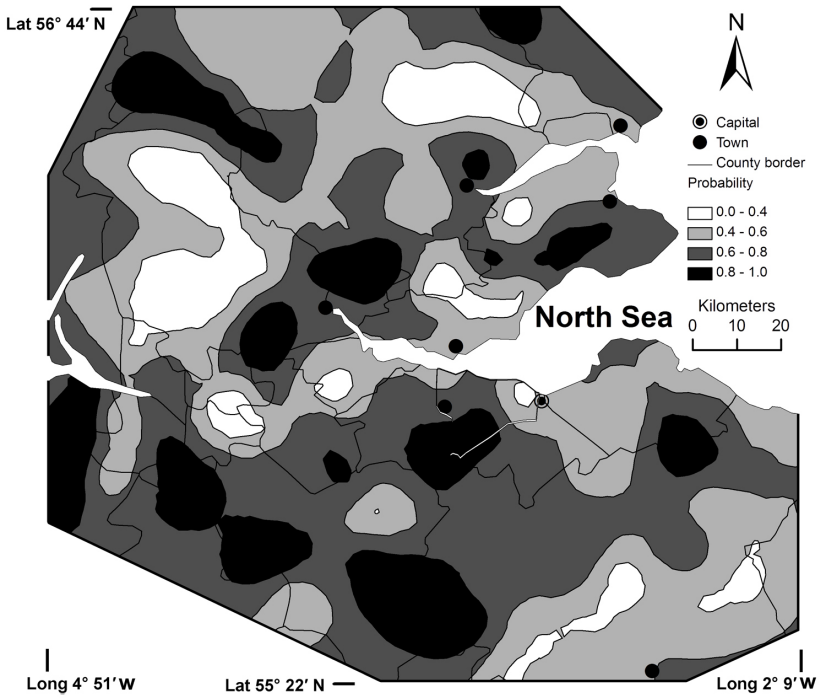


Figure 5.23: Disjunctive kriging for *Engineered* ($> 30\%$).

The maps showing probabilities of exceeding 3 m flooding depth associated with the variables *Mean Flooding Depth* and *Managed Mean Flooding Depth* should be used to estimate the likely return in flood infrastructure investment throughout the study area (Figure 5.24 and Figure 5.25). The higher the probability, the more likely it is that an existing or planned SFRB is making a positive impact on flood control. In contrast to the *Mean Flooding Depth*, the map for *Managed Mean Flooding Depth* indicates

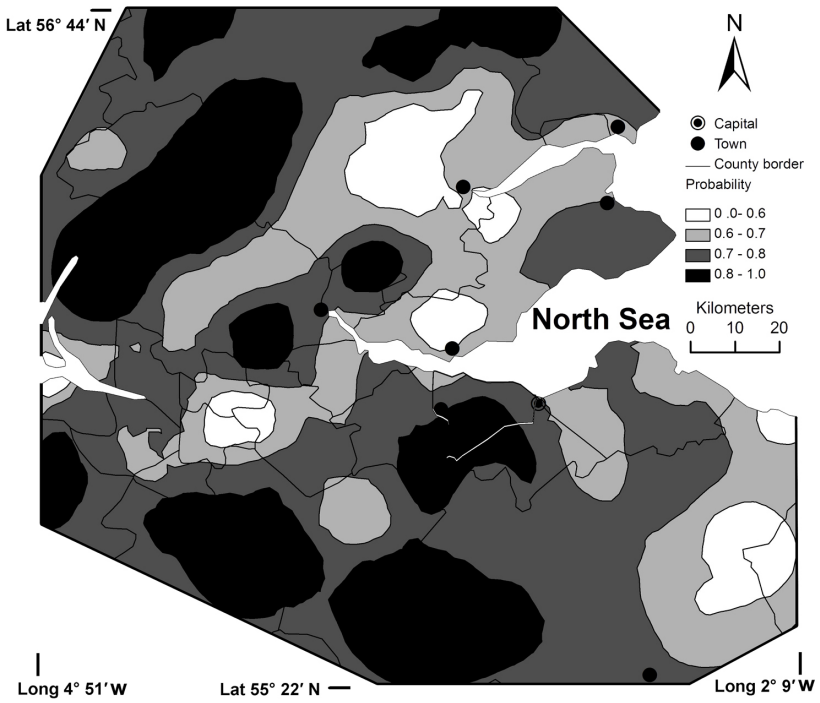


Figure 5.24: Disjunctive kriging for *Mean Flooding Depth* ($> 3m$).

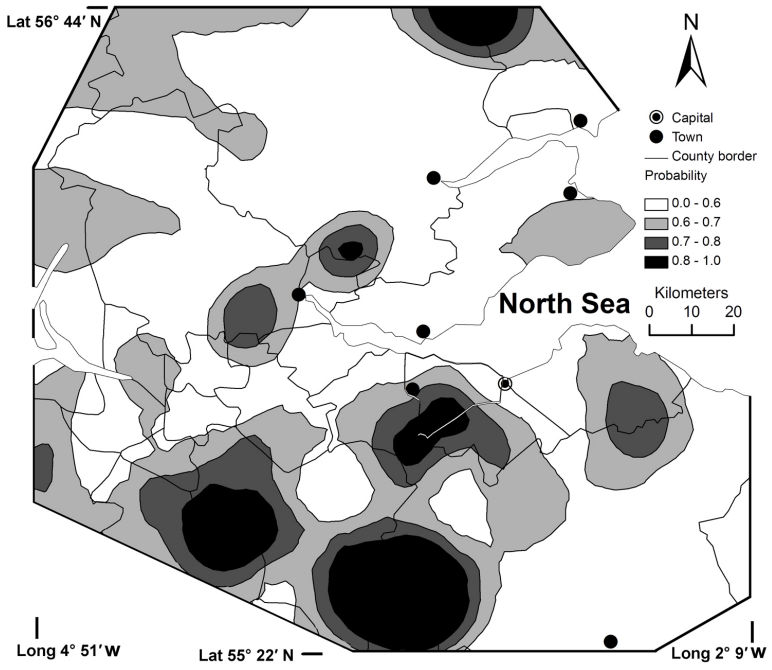


Figure 5.25: Disjunctive kriging for *Managed Mean Flooding Depth* ($> 3m$).

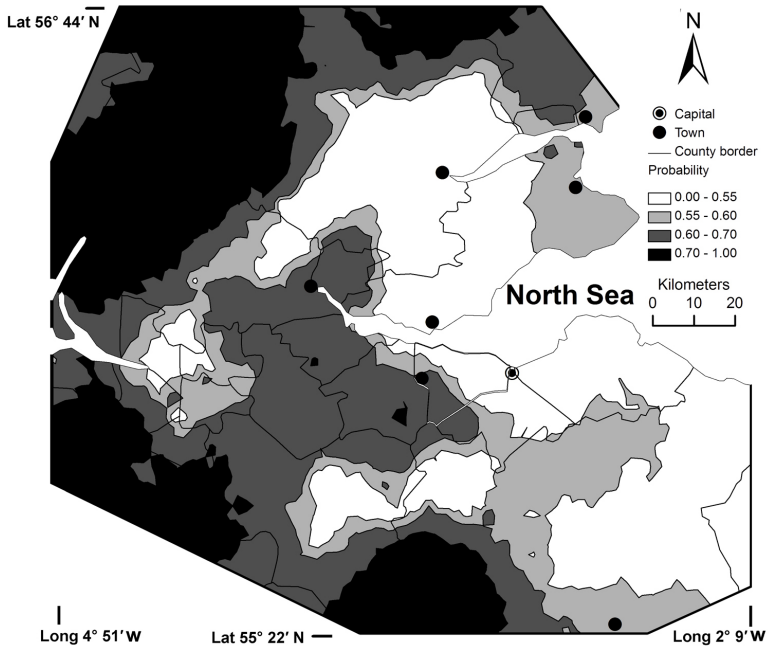


Figure 5.26: Disjunctive kriging for *Maximum Flood Water Volume*.

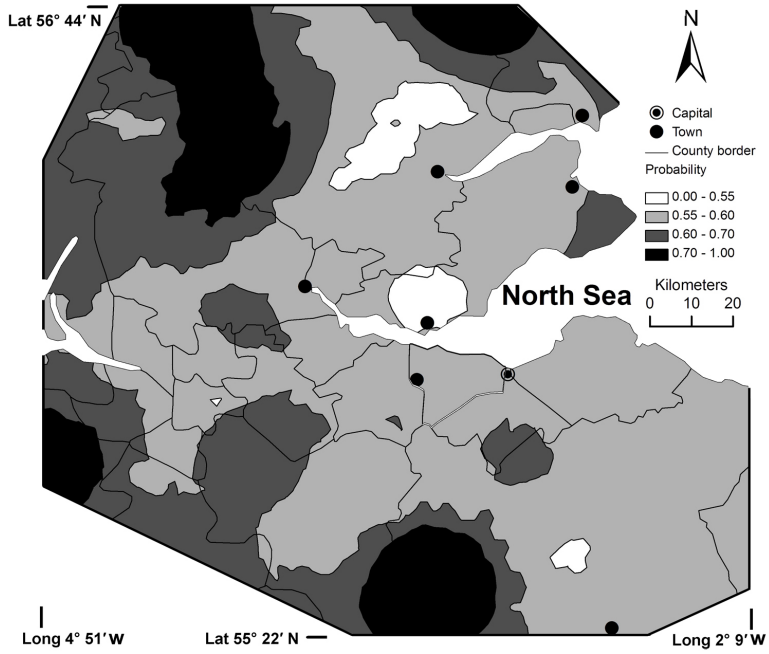


Figure 5.27: Disjunctive kriging for *Managed Maximum Flood Water Volume*.

much lower probability values. Moreover, only those areas that are dominated by reservoirs which could be used by Scottish Water and local authorities for flood control

management are shown. The greatest potential for active flood control is in areas situated to the south-west of the capital such as the Pentland Hills.

Figure 5.26 and Figure 5.27 show that the areas with the greatest flood storage capacity are located in upland catchments distant from populated lowland areas. The probabilities for the likely volumes that could be used for active flood control management by Scottish Water and local authorities are clearly shown in Figure 5.27, because unmanageable storage volumes within natural water bodies have been excluded from the probability map. A comparison of Figure 5.23 and Figure 5.27 indicates that the north-west of the study area has the greatest potential for low-cost SFRB investment yielding a high flood water storage volume return.

5.6.4 Consequences for Flood Risk Management in Scotland

As found in the cluster analysis of the SFRB data (see Section 5.1), Traditional Flood Retention Basins (SFRB type 2) and Natural Flood Retention Wetlands (SFRB type 6) were the two dominant SFRB types in central Scotland area. Figure 5.28 shows an example of a water body classified as a Traditional Flood Retention Basin, however, currently it is only used for drinking water supply purposes. There is great but under utilized potential in using former and less important current potable supply reservoirs for flood control purposes. In comparison, Figure 5.29 is a representative example of a Natural Flood Retention Wetland which is predominantly used for environmental protection and recreational purposes, and has limited flood control potential.

The fieldwork program identified a large number of former water supply reservoirs, which were predominantly identified as SFRB type 2. In the vast majority of cases, these structures now fulfil multiple roles providing opportunities for recreation, nature conservation and angling with many former drinking water reservoirs or industrial water supply structures being managed as fisheries.

Figure 5.28: Glenfarg Reservoir (near to Rossie Ochill, County of Perth and Kinross), an example of a Traditional Flood Retention Basin (type 2) that is, however, currently used only for drinking water supply purposes.

A feature of these sites, based on the majority of current (SFRB type 1) and former (SFRB type 2; Figure 5.28) drinking water supply reservoirs surveyed, is that they are maintained at their maximum water retention volumes, and their corresponding spillways are continuously in operation. In this mode of operation, the extensive infrastructure is making very little contribution to water retention (i.e. flood control) in the upper catchments.

It follows that a change in management practice of reservoir-like structures by Scottish Water and local authorities could assist in sustainable flood risk management planning, leading to more sustainable reservoirs. Effectively, this would require some water to be released from the reservoirs prior to expected heavy precipitation. As the vast majority of former drinking water reservoirs have manual water level control, this would require on-site visits to manually release excess water, only closing the reservoir prior to a heavy rainfall event. This simple operation would enhance the reservoir capacity for water

Figure 5.29: Morton Loch (near Fife), an example of a Natural Flood Retention Wetland (type 6) that is predominantly used for environmental protection, recreational and diffuse pollution control purposes.

storage in the upper reach of the catchment and retard the peak flows from the upper catchment, which is likely to lead to a reduction of flooding downstream. Combining this approach with conventional solutions such as sustainable drainage systems, barriers and dykes will help to reduce the size, cost and land take of other flood defences. It is critical to the success of such an approach that appropriate compensation is provided to the owners of the structures to reflect the value of this service and the mild inconvenience it may cause. However, it is beyond the scope of this thesis to discuss how to evaluate the value of the service in practice. Critical issues to be addressed in this approach are the needs of the owners and operators of the reservoirs. In particular, many of these reservoirs are leased to fishing clubs, typically for Salmo trutta (brown trout) angling. A balance between the requirements of anglers and water quality targets need to be found. The majority of these sites only operate as fisheries between the 15 March and 15 October, which is the traditional trout fishing season in Scotland. As the most severe

rainfall and storm events are predicted for the winter months [34], the reservoirs could be used for flood control purposes outside the fishing season (i.e. shift from SFRB type 5 to SFRB type 3). A major concern of the fisheries owners will be the retention of the fish within the reservoirs during periods of water release, and this may require the fitting of fine screens onto the valve controlled outlets of a reservoir. Equally, water supply organizations such as Scottish Water will need to be reassured that the change of management practice will not impact negatively on the water quality (required for SFRB type 2) within the basin and any management action would need to ensure that all the SFRB purposes and uses are maintained.

The proposed management change can be integrated into spatial planning by justifying the types of SFRB. For example, the SFRB concept could support the Water of Leith Flood Prevention Scheme to protect Edinburgh from flooding [109]. A proper classification of the SFRB located within the Water of Leith catchment area that have flood control potential would clarify their individual planning status. Clarification of their current purpose (e.g. water supply, flood attenuation, recreation and/or environmental protection) would benefit communication between all stakeholders (e.g. local authorities, land owners and Scottish Water) involved with this case study to optimize their planning effort. Moreover, a geo-statistical approach to river network management, already attempted in France [154] and Iran [86], may also benefit flood risk management planning in Scotland.

5.7 Dam Failure Assessment of SFRB

5.7.1 Dam Failure Assessment for Different Types of SFRB in Scotland

Flood risk can be defined as a function of probability of occurrence and extent of damage, the later of which consists of two factors, damage potential and vulnerability

Table 5.13: Summary statistics for key variables relevant for the determination of the risk-related SFRB.

Variables	Type 1 (9 sites)	Type 2 (134 sites)	Type 3 (24 sites)	Type 4 (10 sites)	Type 5 (16 sites)	Type 6 (6 sites)
E	98.6±0.9	70.3±10.4	28.3±8.0	26.5±9.7	32.2±10.5	13±2.7
DH	30.7±18.9	11.5±8.9	2.1±1.0	2.3±0.9	2.4±2.2	0.5±0.9
DL	289.8±172.0	278.3±199	132.5±96.2	68.5±52.2	98.8±107	10±14.1
MFWV	116.3±320.4	3.1±6.9	0.1±0.2	0.1±0	0.9±1.6	16.4±25.1
MAR	10.0±36.8	107.1±28.7	114.7±30.2	96.9±26.2	94.0±23.3	88.4±14.8
CZ	129.8±189.3	7.9±11	1.7±1.5	4.3±5.9	11±26.7	39.6±71.5
MDB	14.9±14.6	6.3±3.7	2.4±0.7	2.4±0.8	2.8±0.9	6.6±9.0
DC	91.8±4.9	78.5±7.5	58.2±12.4	47.8±19.9	57±22.1	62.2±17.8
DFH	12.6±12.5	10±13	2.8±3.3	1.2±1.8	4.3±5.1	1.8±2.3
DFR	6.4±3.0	6.2±3.8	4.5±3.9	4.4±3.1	5.1±3.6	3.7±4.4

Note: E (*Engineered*, %); DH (*Dam Height*, m); DL (*Dam Length*, m); MFWV (*Maximum Flood Water Volume*, $millionm^3$); MAR (*Mean Annual Rainfall*, cm/a); CZ (*Catchment Size*, km^2); MDB (*Mean Depth of Basin*, m); DC (*Dam Condition*, %); DFH (*Dam Failure Hazard*, %); DFR (*Dam Failure Risk*, %).

[171, 156, 89]. Furthermore, risk assessments may also take relevant geographical and statistical data into account [171]. All these factors have been incorporated by the proposed variable *Dam Failure Risk*.

The number of potentially affected inhabitants and business activities as well as environmental damage has been addressed by the proposed new variables *Dam Failure Hazard*. To reflect the levels of the hazards and risks of dam failure, the vulnerability indicators were assumed to vary between low, moderate, high and very high values.

This project combined the above variables with risk management and assessment to make a rapid tool. The 199 surveyed SFRB consist of 6 types, mainly of which belong

to Type 2 (134 sites). Table 5.13 shows the summary statistics for the three risk-related variables as well as their relevant key variables. It indicates that different types of SFRB are associated with different levels of hazards and risks of dam failure. For instance, SFRB of Type 1 (9 sites) had the highest *Dam Failure Hazard* (12.6%, high level) and *Dam Failure Risk* (6.4%, high level). These basins had the largest mean values for the variables of *Engineered*, *Dam Height*, *Dam Length*, *Depth of Basin*, *Flood Water Volume and Catchment Size*, which would place significant influence on dam failure hazards and risks. Since the SFRB in Type 1 are usually used for hydraulic electric stations, they are well maintained and thus had the highest *Dam Condition* (91.8%). Following Type 1, the SFRB of Type 2, which mainly comprises Scottish drinking water reservoirs, having relative large *Dam Height*, *Dam Length*, *Maximum Flood Water Volume* and *Catchment Size*, were often associated with relative high hazard (10%, moderate level) and risk (6.2%, high level) of dam failure. Most of the SFRB in Type 2 are reservoirs and are managed by agencies or authorities, therefore, their *Dam Condition* (78.5%) was relatively high.

Basins of SFRB types 3 and 4 are often in small size and with low engineered structures, therefore, the *Dam Failure Hazard* and *Dam Failure Risk* were both low. Type 6 of SFRB had low *Dam Failure Hazard* (1.8%, low level) and lowest *Dam Failure Risk* (3.7%, low level). These basins mainly are lochs and rivers having relative large *Flood Water Volume*, *Catchement Size* and *Depth of Basin* but with lowest mean values for *Engineered*, *Dam Height* and *Dam Length*. In contrast, SFRB in Type 5 had slightly higher dam failure hazards and risks than that of types 3, 4 and 6. It might because these basins of Type 5 mainly are located near residence and used for public parks, recreations and water sport.

5.7.2 Spatial Distribution of the Dam Failure Hazards and Risks

To illustrate the spatial characters of the levels of dam failure hazard and risk across the research area, Ordinary Kriging was applied. Figures 5.30 and 5.31 show the spatial distribution maps based on interpolations of the *Dam Failure Hazard* and *Dam Failure Risk* for SFRB in Scotland, respectively. Figure 5.30 indicates that the *Dam Failure Hazard* for Scottish SFRB varies from 0 to 66%. A large part of the research area was at the moderate level of *Dam Failure Hazard* (between 5% and 10%), such as the areas around Edinburgh, Livingston, Stirling and Dunfermline. However, the *Dam Failure Hazard* near the city of Glasgow and the Northeast of Stirling was relative high between 10% and 15% (high level). A limited number of sites were predicted with very high *Dam Failure Hazard*. For such areas, the managers should pay much attention to the SFRB dam safety.

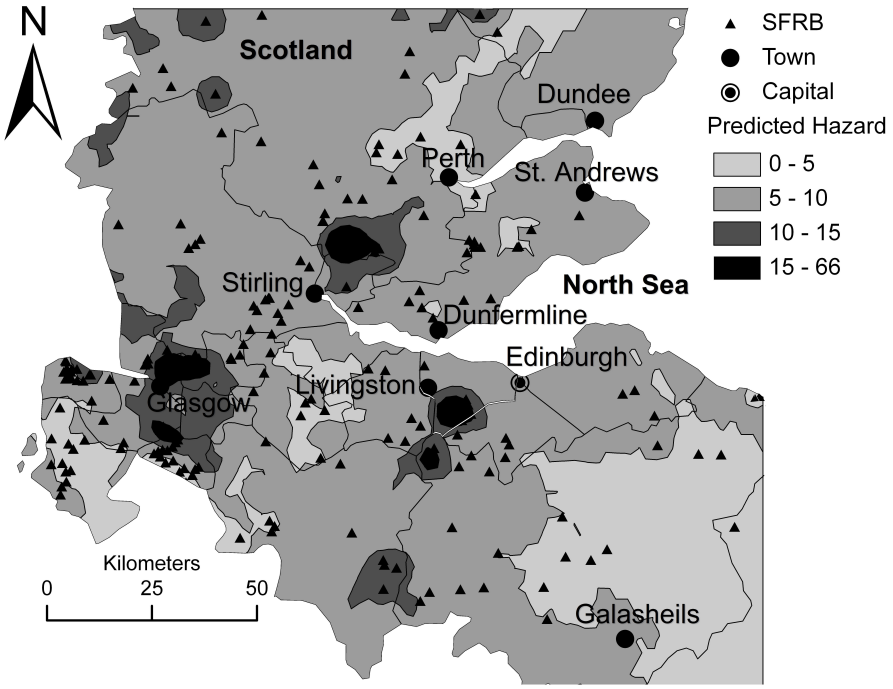


Figure 5.30: Spatial distribution of the *Dam Failure Hazard* for Sustainable Flood Retention Basins in central Scotland.

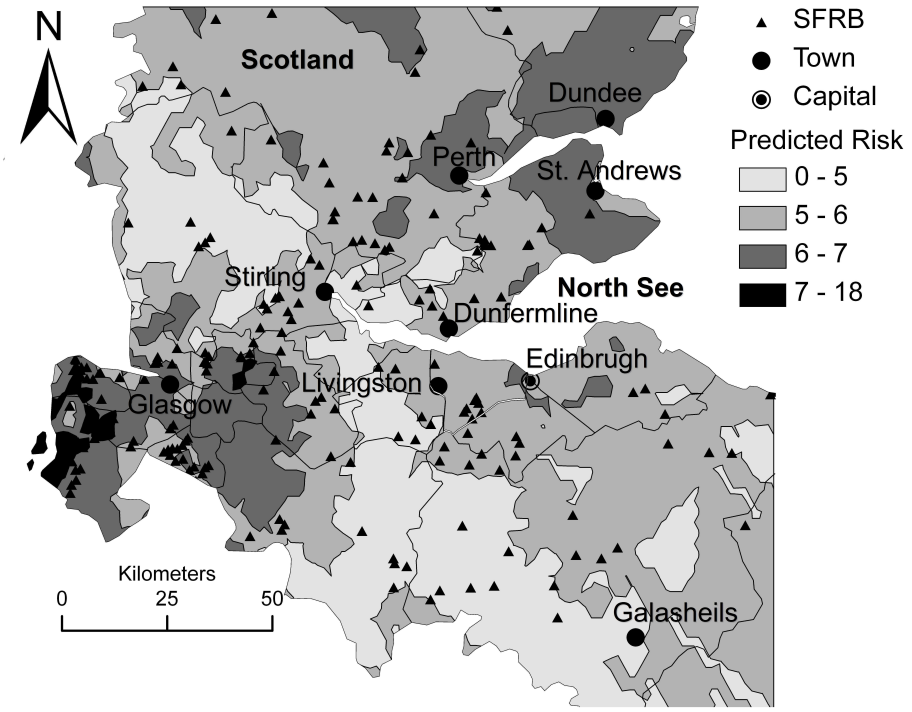


Figure 5.31: Spatial distribution of the *Dam Failure Risk* for Sustainable Flood Retention Basins in central Scotland.

Figure 5.31 shows that SFRB situated at south of Livingston and northwest of Stirling have low *Dam Failure Risk* ranging from 0 to 5%. SFRB located around Edinburgh, Livingston, Dunfermline and Stirling had higher risk of between 5% and 6% (moderate level). Most of the research area was covered by the above two ranges. While the SFRB near Glasgow and Perth had high level of *Dam Failure Risk*, which is between 6% and 7%. Some sites that have even higher *Flood Failure Risk* are located at the west of Glasgow. In addition, it is noticeable that the *Dam Failure Risk* of SFRB located near the towns were higher than that located in relatively remote areas.

The above findings provide the decision-makers or planners with spatial support for flood risk management. By considering distribution of the levels of hazard and risk, they could make specific plans or strategies for different regions to lower the hazard and risk resulted from dam failure.

5.7.3 Risk categories

The dam safety community is discussing the categorization of reservoirs according to simple risk-related criteria for legislative purposes. Table 5.14 shows the overview of the risk category of the surveyed SFRB with respect to *Dam Height*, *Maximum Flood Water Volume*, *Loss of Life Risk* and infrastructure damage. The bin borders have been selected according to recent governmental consultation discussions predominantly in the UK and Germany. For example, the first three risk categories agree with those currently proposed by the Scottish Government [109].

Table 5.14 indicates that most SFRB sites in Scotland are in the low and moderate risk categories. This is consistent with the spatial distribution as illustrated on figure 5.31. Precisely 88 SFRB were considered to be low risk sites, while 99 SFRB belonged to the moderate risk category. Only 3 SFRB were assigned to the high risk category, all of which had relative high dams of more than 5 m. Seven SFRB were associated with a very high risk. This might be because these sites have both high dams and *Maximum Flood Water Volume* ($> 10^5 m^3$), and are frequently located in densely populated areas.

5.8 Summary

In this chapter, experimental results and discussions were demonstrated. Specifically, the meaningful clusters were identified and matched with 6 SFRB types. The dimensionality of SFRB data was reduced into a lower dimensional space by PCA. The correlations among 43 variables were visualized via SOM mapping and the variables and SFRB types were predicted with SOM model. The feature selection succeeded in selecting small subset of variables to achieve high classification accuracy. Multi-label classification further revealed the complex SFRB functions and statuses. Spatial analysis of SFRB provided reasonable suggestions for spatial planning of flood control. Dam failure assessment of SFRB provided a rapid tool for preliminary evaluation of the flood risk.

Table 5.14: Overview of sustainable flood retention basins (SFRB) with respect to *Dam Height*, *Maximum Flood Water Volume*, *Loss of Life Risk* and *Risk of Infrastructure Failure* (damage). The first three risk categories are equivalent to Scottish definitions.

Death and Damage	Maximum Flood Water Volume (m^3)	SFRB in Scotland			
		Dam Height			
		No dam	$< 5m$	5 to $15m$	$> 15m$
1. Low risk (minor risk of damage to property downstream)					
< 1 person dies and minor damage	$<$ 10k	0	0	0	0
	$\geq$10k to 25k	0	2	1	0
	$\geq$25k to 100k	1	12	2	0
	$\geq$100k	5	19	34	12
2. Moderate risk (moderate risk to damage to property and infrastructure downstream)					
< 1 person dies and moderate damage	$<$ 10k	0	1	0	0
	$\geq$10k to 25k	0	2	0	0
	$\geq$25k to 100k	0	12	2	0
	$\geq$100k	1	14	47	20
3. High risk (risk to life and/or significant risk to property and critical infrastructure)					
≥ 1 to ≤ 20 people die and high damage	$<$ 10k	0	0	0	0
	$\geq$10k to 25k	0	0	0	0
	$\geq$25k to 100k	0	0	0	0
	$\geq$100k	0	0	3	0
4. Very high risk (high risk to life and significant risk to property and critical infrastructure)					
≥ 20 people die and high damage	$<$ 10k	0	0	0	0
	$\geq$10k to 25k	0	0	0	0
	$\geq$25k to 100k	0	0	0	0
	$\geq$100k	0	1	5	1

Note: k=1000.

Chapter 6

CONCLUSIONS

In this thesis, I focus on the rapid survey method of Sustainable Flood Retention Basins (SFRB) and the comprehensive analysis of SFRB by exploring a wide variety of algorithms. Representative cases studies were further investigated to verify the efficiency of the proposed models or frameworks. This chapter starts to summarize the major achievements of this thesis in Section 6.1 and then addresses the research limitations in Section 6.2 and finally points out possible directions for future research in Section 6.3.

6.1 Contributions

1. For SFRB investigation, a guidance manual was established for rapid and comprehensive survey of SFRB. It illustrated the philosophy behind the guidance manual, purposes of SFRB, detailed explanations to the characteristic variables, the criteria on estimating values for them, and the field survey templates. Therefore, the guidance manual aids users to understand SFRB quickly and then learn how to undertake SFRB assessment easily and consistently, with different participants or at different times. Besides, the guidance manual can also be used

to identify potential SFRB. As a whole, the guidance manual would serve as a handbook or benchmark for people who are interested in SFRB study.

2. To investigate the intrinsic structure of the SFRB data set, four agglomerative hierarchical clustering algorithms (Single Link, Average Link, Complete Link and Ward's Link) were applied. The results showed that Ward's Link obtained the best results. For Ward's Link, meaningful groups of SFRB could be visualized and discovered through the cluster tree (also called dendrogram). For Scottish SFRB data set, it was obvious that two groups existed (cf. Figure 5.4). One group mainly included SFRB type 2 and type 1 while another group comprised of types 3, 4, 5 and 6 respectively. This is in line with the fact that these two large groups are significantly different from each other. With the splitting of the dendrogram at a lower level, 12 main clusters could be identified which mainly corresponding to the six types of SFRB with high precision and recall (cf. Section 5.1). In particular, type 2 and type 6 were the two largest groups of SFRB. These findings indicate that clustering based on the Ward criterion is efficient and effective to reveal the intrinsic structures hidden in the SFRB data set. However, different clustering algorithms might be appropriate for different data patterns.

3. For a holistic analysis of SFRB, 43 variables were collected to characterize the property of each SFRB. However, some variables may correlate and result in redundancy of the data set. Principal component analysis (PCA) was thus used to analyze the redundancy and identify the relatively important variables of the SFRB data. The PCA transformed the SFRB data from the original 43 dimensions to a new 23-dimensional space with little information lost. This indicated that the SFRB data could be simplified with less variables, which was further explained by feature selection. Since the principal components decomposed by PCA were the linear combinations of all the original variables, it is non-trivial task to interpret the results directly. To analyze the independent and important variables for SFRB characterization, the contribution of each original variable to all 23 principal components was computed by analyzing the eigenvalues

and eigenvectors generated during PCA. Results showed that the contributions of the variables for principal components were slightly different, ranging from 5.69 to 3.81, which indicated that all the variables (except for variable 34) were relatively important for characterizing the SFRB data. In general, PCA can analyze the redundancy of the SFRB data to some extent, however, it is difficult to interpret the new features (principal component) and identify the key variables of SFRB.

4. To comprehensively assess the SFRB, all the 43 variables should be collected during the site investigation. However, some variables of SFRB were difficult or expensive to obtain. To predict the missing values of these variables was a challenge for SFRB analysis. In this thesis, Self-organizing Map (SOM) was introduced to analyze the data structure and predict the missing values of variables as well as the possible types of SFRB. On one hand, SOM brought a deep insight into SFRB data patterns by vector quantization. In comparison with dendrogram in hierarchal clustering, SOM provided more intuitive way to visualize the data structure and show the relationships among the 43 variables. On the other hand, through competitive learning, the SFRB with similar attributes moved together and the missing values of variables can be effectively estimated by finding the best matching unit (BMU). To better estimate the missing value of variable, the variables which were closely related to the predicted variable were selected by computing the Pearson correlation coefficient. Experimental results indicated that SOM can well predict the SFRB variables. Similarly, treating SFRB type as a variable, SOM can also be used to predict SFRB types. Types 1, 2 and 6 of SFRB were predicted well while it was difficult to distinguish SFRB types 3, 4 and 5.

5. Like PCA, the feature selection techniques were designed to identify the most relevant characteristic variables and remove any redundant variables. It indicated that Mutual Information, Information Gain and Relief performed well in identifying and ranking the most relevant SFRB classification variables. However, different feature selection algorithms generated different results regarding the importance of the variables due to their different ranking strategies. A final list of

priority variables was obtained by comparing and combining the feature selection methods with each other. Finally, nine variables were regarded as the most important ones for characterizing and classifying a SFRB. In addition, in order to evaluate the selected features, four classifiers SVM, KNN, NB and J48 were applied to classify SFRB data by using different numbers of variables. The results indicated that the selected nine variables were sufficient for the four classifiers to achieve high classification accuracy. In contrast, the classification accuracy even deceased for the KNN classifier with the number of variables ranging from 10 to 40. This indicated that introducing more variables might lead to more redundancy or dependency. Therefore, identification of the most relevant key variables played a significant role in removing the redundant information, which allowed SFRB assessment to be cost and time efficient. With six typical case studies, it has been further verified that the selected nine important variables had very different performances on individual SFRB types. In comparison with PCA and SOM, feature selection techniques provide a more efficient and effective way to identify the most important characteristic variables of SFRB.

6. Traditional classification schemes have been developed to distinguish SFRB types. However, these methods were limited to assigning only one label to one SFRB site, which cannot reflect what has been observed in the real world. Actually, one basin often has multiple functions and should belong to more than one SFRB type. Therefore, to better understand, assess and manage SFRB, three multi-label classification algorithms (MLSVM, MLKNN and MLBP) were introduced to predict the SFRB types automatically. The experiments showed that all the three multi-label classifiers achieved good results and outperformed the corresponding traditional classifiers (SVM, KNN and BP). In addition, three SFRB case studies were investigated further to verify the effectiveness and benefits of the multi-label learning algorithms. The findings showed that the predicted multiple types would bring a deeper and more comprehensive insight of the status and functions of SFRB than the traditional classification methods. The predicted multiple functions of SFRB thus help to reduce and/or avoid confusions and misunderstandings concerning SFRB assessment and management among

planners, engineers and authorities. Moreover, it helps designers to integrate aesthetics and recreation into SFRB design making SFRB more sustainable and diverse. However, the proposed multi-label classification framework does not take account of uncertainty issues.

7. To analyze the spatial property of SFRB data, the geo-statistical analysis techniques of ordinary kriging and disjunctive kriging were applied to SFRB management in the context of Floor Risk Management Plans. Five flood-related variables were focused: *Engineered*, *Mean Flooding Depth*, *Managed Mean Flooding Depth*, *Maximum Flood Water Volume*, and *Managed Maximum Flood Water Volume*. The ordinary kriging allowed for a clear interpretation of areas requiring further flood control investment while disjunctive kriging was used to assess the probability of individual variables to exceed specific management thresholds related to flood control. With the help of ordinary kriging maps, decision makers could intuitively identify the flood prone areas and undertake spatial planning. While the disjunctive kriging maps are very useful references for designers when they design new SFRB since it provides the possibility that the true value exceeds the threshold. Depending on different needs in practice, different techniques can be preferably or comprehensively used. Furthermore, the proposed geo-statistical methodology would aid stakeholder communication by delivering information regarding the most favorable locations for SFRB investment.

8. Dams are innately hazardous structures. To assess the dam failure hazards and risks of SFRB, three new risk-relate variables (*Dam Condition*, *Dam Failure Hazard* and *Dam Failure Risk*) and a rapid screening tool was proposed based on expert judgement. Findings showed that different SFRB types had different dam failure hazard and risk levels. In the context of the Flood Directive, to integrate spatial planning into flood risk management, Ordinary Kriging was applied to map the spatial distributions of *Dam Failure Hazard* and *Dam Failure Risk* across the research area. The results demonstrated that the hazards and risks of dam failure varied in different regions of central Scotland, which could help decision

makers to intuitively identify the risk areas requiring emergent coping measures. However, the detailed dam failure risks and hazards associated with these areas need to be further assessed and managed.

6.2 Research limitations and Recommendation

In this thesis, we provided a series of techniques to analyze the SFRB data. However, there also exist some limitations, which are listed as below.

1. The guidance manual for the assessment of water bodies including SFRB is one of a number of tools that are being produced by the Interreg IVB project of Strategic Alliance for integrated Water Management Actions (SAWA). Each tool is applicable to different situations and to a range of different circumstances. Moreover, the guidance manual is suitable to limited research areas but not across the whole world. It is recommended that the most appropriate tool should be selected for a specific situation.

2. The dam failure assessment approach presented has its clear limitations. For example, the proposed composite *Dam Failure Risk* variable has shortcomings. The flood risk is a complex issue affected by many visible and invisible factors. It needs more informative support for the estimations.

3. The proposed frameworks or tools in this thesis, which have been designed for water bodies and flood defence structures in Southern Baden and central Scotland, can be extended and applied to other regions in the world with temperate or oceanic climate by stakeholders and environmental engineering scientists for decision-making processes. However, local and regional variables such as *Seasonal Influence*, *Typical Wetness Duration* and *Drainage* might require adjustment by experts for practical cases. Moreover, each method has its own advantages and disadvantages, dedicated to resolve certain specific problems. Therefore, users should choose an appropriate method according to the real

problems at hand. For instance, to identify the key characteristic variables of SFRB, feature selection techniques are recommended rather than PCA or SOM, since their results are more effective and easier to interpret. Moreover, further recommendation is to find all the potential important variables in combination with the classification performances. In contrast, to check how redundant one data set is, PCA can be applied. But if the purpose is to visualize the correlations among SFRB variables and/or to predict the difficult-to-determine variables of SFRB , SOM is recommended. Taking the SFRB design as an example, it is suggested that planners should take multi-label classification into account, making the new SFRB more sustainable and have diverse functions. To make spatial planning for flood risk management, Ordinary kriging is recommended for decision-makers since it intuitively provides the risk zones while Disjunctive kriging is more valuable for SFRB designers since it shows the estimated probability that the value of a specific SFRB variable exceeds one threshold.

4. In this study, data was acquired from various resources such as research papers, books, digital maps, archives, agencies, institutions, governments and the experts' experiences which also come from the accumulation of knowledge. It means that the data quality is significantly impacted by the availability of the required information. Moreover, to avoid or reduce the assessors' bias toward the SFRB survey due to their disciplines, it is recommended that the survey team may consists of interdisciplinary researchers. Furthermore, the team members should assess the same SFRB independently and then discuss together to achieve a united report.

5. It is recommended that the SFRB can be designed, maintained and managed sustainably to adapt to climate change across Europe. In Southern Baden, most SFRB were purpose built used for flood protection. Additionally, it was observed in summer 2010 that lots of existing SFRB were being upgraded and many new SFRB were being built up in the context of climate change. In contrast, in central Scotland, only one SFRB was noticed being upgraded for flood defense. Actually, the SFRB in central Scotland have high potential to be developed to contribute

to flood control. Similarly, adapting SFRB to climate change can be extended to other European countries.

6.3 Outlook

The current research has produced a detailed guidance manual for SFRB survey and the findings have demonstrated the effectiveness and efficiency of the proposed frameworks for comprehensive analysis and assessment of SFRB. Nevertheless, obviously, there is still a requirement to carry out further studies. Some important future research frontiers are as follows.

1. **Updating Guidance Manual**: With more knowledge to Sustainable Flood Retention Basins (SFRB), more new variables may be proposed to supplement the current 43 variables to better characterize SFRB. Therefore, accordingly, the guidance manual would need to be expanded and updated.

2. **Subspace Clustering**: The clustering of SFRB in the current study has helped to identify the intrinsic groups of SFRB based on the whole set of 43 variables. However, there might be some clusters which only exist on subspace dimensions rather than on the whole 43 dimensions, which is known to be a subspace clustering problem. Exploring the subspace clustering of SFRB will be useful to better understand the structural patterns of the SFRB data.

3. **Uncertainty Exploration**: In the real world, data are always contaminated due to environmental factors, measurement precision, and human factors, etc. Regarding SFRB, the value for each characteristic variable was associated with uncertainty due to estimation or measurement error. The current frameworks did not take account of uncertainty issues. Therefore, in order to better represent the inherent property of each variable, a model adapted to cope with the uncertainty of SFRB data should be explored. Furthermore, based on the modeled uncertainty of the SFRB data, uncertain clustering and classification

can be explored. In this case, the mining of the uncertain SFRB data will reflect the reality more truthfully and it will help to gain deeper insights into SFRB.

4. **Flood Risk Assessment System**: It would be very worthwhile to integrate the uncertainty into the establishment of a real-time flood risk assessment system. Meanwhile, the association rules between the flood risks and other SFRB variables need to be exploited. It is supposed to produce a reliable and efficient risk assessment system which allows updating associate rules and predicting risks of basins automatically. In that case, it will dramatically improve the implementation of the Flood Directive for Flood Risk Management Plan.

References

[1] Aha, D. W., Kibler, D., and Albert, M. K. Instance-based learning algorithms. *Journal of Coastal Research 6*, 1 (1991), 37–66.

[2] Alhoniemi, E., Hollmn, J., Simula, O., and Vesanto, J. Process monitoring and modeling using the self-organizing map. *Integrated Computer Aided Engineering 6*, 1 (1999), 3–14.

[3] Almuallim, H., and Dietterich, T. G. Learning with many irrelevant features. In *Proceedings of the 9th National Conference on Artificial Intelligence* (Anaheim, CA, 1991), AAAI Press, pp. 547–552.

[4] Anderson, D. M., Gilbert, P. M., and Burkholder, J. M. Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. *Estuaries 25*, 4B (2002), 704–726.

[5] Andjelkovic, I. Guidelines on non-structural measures in urban flood management. Tech. Rep. Technical Documents in Hydrology 50, UNESCO, Paris, 2001.

[6] Astela, A., Tsakovskib, S., Barbieric, P., and Simeonovb, V. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Research 41*, 19 (2007), 4566–4578.

[7] ate V. Heal, K. Manganese and land-use in upland catchments in scotland. *The Science of The Total Environment 265*, 1-3 (2001), 169–179.

[8] Baggaley, N., Langan, S., Futter, M., Potts, J., and Dunn, S. Long-term trends in hydro-climatology of a major scottish mountain river. *Science of the Total Environment 407*, 16 (2009), 4633–4641.

[9] Bah, A. R., Kravchuk, O., and Kirchhof, G. Sensitivity of drainage to rainfall, vegetation and soil characteristics. *Computers and Electronics in Agriculture 68*, 1 (2009), 1–8.

[10] Bassett, D., Pettit, A., Anderton, C., and Grace, P. Scottish flood defence asset database - final report. Tech. rep., JBA Consulting for the Scottish

Government, Aug. 2007. http://www.scotland.gov.uk/Publications/2007/08/20111904/0.

[11] Bayley, S. E., and Guimond, J. K. Effects of river connectivity on marsh vegetation community structure and species richness in montane floodplain wetlands in jasper national park, alberta, canada. *Ecoscience 15*, 3 (2008), 377–388.

[12] Berkhin, P. Probabilistic principal component analysis. *Techniques 10* (2002), 1–56.

[13] Borak, J. S., and Strahler, A. H. Feature selection and land cover classification of a modis-like data set for a semiarid environment. *International Journal of Remote Sensing 20*, 5 (1999), 919–938.

[14] Boutell, M. R., Luo, J., Shen, X., and Brown, C. M. Learning multi-label scene classification. *Pattern Recognition 37*, 9 (2004), 1757–1771.

[15] Bowles, D. S., Anderson, L. R., and Glover, T. F. Comparison of hazard criteria with acceptable risk criteria. In *Proceedings of the Annual Meeting of the Association of State Dam Safety Officials* (Atlanta, Georgia, 1995), pp. 293–302.

[16] Bowles1, D. S., Anderson, L. R., Glover, T. F., and Chauhan, S. S. Portfolio risk assessment: A tool or dam safety risk management. In *Proceedings of the 1998 USCOLD Annual Lecture* (1998), Buffalo, New York.

[17] Bronstert, A., Baardossy, A., Bismuth, C., Buiteveld, H., Disse, M., Engel, H., Fritsch, U., Hundecha, Y., Lammersen, R., Niehoff, D., and Ritter, N. Multi-scale modelling of land-use change and river training effects on floods in the rhine basin. *River Research and Applications 23*, 10 (2007), 1102–1125.

[18] Brown, R. Soil texture. Tech. Rep. Fact Sheet SL-29, University of Florida, Institute of Food and Agricultural Sciences Extension, Sep. 2003.

[19] Byrne, R. Classified list of legistration in ireland. Tech. rep., Law Reform Commission, Dublin, Dec. 2010. http://www.lawreform.ie/_fileupload/consultation%20papers/full.pdf.

[20] Casali, J., Gimenez, R., Santisteban, L. D., Alvarez-Mozos, J., Mena, J., and de Lersundi, J. D. V. Determination of long-term erosion rates in vineyards of navarre (spain) using botanical benchmarks. *Catena 78*, 1 (2009), 12–19.

[21] CEH. Flood estimation handbook. Tech. Rep. Fact Sheet SL-29, Center for Ecology and Hydrology, Wallingford, Egland, UK, 1999. http://www.ceh.ac.uk/Feh2/InsidetheFloodEstimationHandbook.html.

[22] Chen, Y.-W., and Lin, C.-J. Combining svms with various feature selection strategies. In *Feature extraction, foundations and applications*, I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Eds.

[23] Chung, E. G., Bombardellia, F. A., and Schladow, S. G. Modeling linkages between sediment resuspension and water quality in a shallow, eutrophic, wind-exposed lake. *Ecological Modelling 220*, 9-10 (2009), 1251–1265.

[24] Ciscar, J.-C. Climate change impacts in europe final report of the peseta research project. Tech. rep., Joint Research Centre, Institute for Prospective Technological Studies, Luxembourg: Publications Office of the European Union, 2009. `http://ftp.jrc.es/EURdoc/JRC55391.pdf`.

[25] Clare, A., and King, R. D. Knowledge discovery in multi-label phenotype data. In *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2001)* (2001), Springer, pp. 42–53.

[26] Cline, M. G. Basic principles of soil classification. *Soil Science 67* (1949), 81C–91.

[27] Colin, F., Puech, C., and de Marsily, G. Relations between triazine flux, catchment topography and distance between maize fields and the drainage network. *Journal of Hydrology 236*, 3-4 (2000), 139–152.

[28] Collins, A. L., and Walling, D. E. The storage and provenance of fine sediment on the channel bed of two contrasting lowland permeable catchments. *River Research and Applications 23*, 4 (2007), 429–450.

[29] Colofon. The quality of drinking water in the european union - synthesis report on the quality of drinking water in the european union period 2002-2004 (directives 80/778/eec and 98/83/ec). Tech. rep., Colofon, 2007. `http://circa.europa.eu/Public/irc/env/drinking_water_rev/library?l=/drinking_synthesis/report_2002-2004pdf/_EN_1.0_&a=d`.

[30] Dash, M., and Liu, H. Feature selection for classification. *Intelligent Data Analysis 1* (1997), 131–156.

[31] Dawson, J. J. C., Soulsby, C., Tetzlaff, D., Hrachowitz, M., Dunn, S. M., and Malcolm, I. A. Influence of hydrology and seasonality on doc exports from three contrasting upland catchments. *Biogeochemistry 90*, 1 (2008), 93–113.

[32] de Bruijn, K. M. Resilience indicators for flood risk management systems of lowland rivers. *International Journal of River Basin Management 2*, 3 (2004), 199–210.

[33] EA. River basin management planning, 2009. EA: Environment Agency.

[34] EAC. Adapting to climate change. Tech. Rep. HC 113, Environmental Audit Committee, London, UK, 2010. `http://www.publications.parliament.uk/pa/cm200910/cmselect/cmenvaud/113/113.pdf`.

[35] EDINA. Digimap and historic digimap collections, 2009. EDINA is a Joint Information Systems Committee (JISC) Data centre based at the University of Edinburgh.

[36] Edwardsa, A. C., and Withers, P. J. A. Transport and delivery of suspended solids, nitrogen and phosphorus from various sources to freshwaters in the uk. *Journal of Hydrology 350*, 3-4 (2008), 144–153.

[37] Elisseeff, A., and Weston, J. A kernel method for multi-labelled classification. In *In Advances in Neural Information Processing Systems 14* (2001), vol. 14, MIT Press, pp. 681–687.

[38] EPC. Directive 2000/60/ec of the european parliament and of the council establishing a framework for the community action in the field of water policy. *Official Journal of the European Union*, Ref L 327 (Dec. 2000), 1–72. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2000:327:0001:0072:EN:PDF.

[39] EPC. Directive 2007/60/ec of the european parliament and of the council of 23 october 2007 on the assessment and management of flood risks. *Official Journal of the European Union*, Ref L 288 (Nov. 2007), 27–34. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:288:0027:0034:EN:PDF.

[40] Estévez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks 20*, 2 (2009), 189–201.

[41] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* (1996), AAAI Press, pp. 226–231.

[42] Evans, S. Y., and Hohl, A. Reservoir inundation mapping and emergency planning. In *Proceedings of Water and Environment 2010* (2010), London, England, UK.

[43] Few, R. Flooding, vulnerability and coping strategies: local responses to a global threat. *Progress in development study 3*, 1 (2003), 43–58.

[44] Feyen, L., Dankers, R., and Boodis, K. Evaluating the benefits of adapting to changing flood hazard in europe. In *IOP Conf. Series: Earth and Environmental Science 6* (2009), vol. 6, IOP Publishing.

[45] Fodor, I. K. A survey of dimension reduction techniques. Tech. rep., Lawrence Livermore National Laboratory, Livermore, CA, 2002. https://computation.llnl.gov/casc/sapphire/pubs/148494.pdf.

[46] Fokkens, B. The dutch strategy for safety and river flood prevention. In *Proceedings of the NATO Advanced Research Workshop on Extreme Hydrological Events: New Concepts for Security* (2007), vol. 8, pp. 337–352.

[47] for Water Research (FWR), F. River basin district, 2009.

[48] Forman, G. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research 3* (2003), 1289=–1305.

[49] Foster, G., Chiverrell, R. C., Harvey, A. M., and Dearing, J. A. Catchment hydro-geomorphological responses to environmental change in the southern uplands of scotland. *Holocene 18*, 6 (2008), 935–950.

[50] Gallardo, A. Spatial variability of soil properties in a floodplain forest in northwest spain. *Ecosystems 6* (2003), 564–576.

[51] Gersho, A., and Gray, R. M. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, Norwell, MA, 1992.

[52] Gilvear, D. J., and Black, A. R. Flood-induced embankment failures on the river tay: implications of climatically induced hydrological change in scotland. *Hydrological Sciences Joural 44*, 3 (1999), 345–362.

[53] Gormley, J., and Mansergh, M. The planning system and flood risk management - consultation draft guidelines for planning authorities. Tech. rep., Environment, Heritage and Local Government, 2008. `http://www.environ.ie/en/Publications/DevelopmentandHousing/Planning/FileDownLoad,18428,en.pdf`.

[54] GPS. Magellan geographical positioning system (gps) receivers, 2009.

[55] Grad, H. Note on n-dimensional hermite polynomials. *Communications on Pure and Applied Mathematics 2*, 4 (2006), 325–330.

[56] Green, C. H., Parker, D. J., and Tunstall, S. M. Assessment of flood control and management options. Tech. rep., World Commission on Dams, Vlaeberg, Cape Town, South Africa, 2000. `http://www.swissdams.ch/Committee/Dossiers/wcd/Thematic%20review/tr44_finaldraft.pdf`.

[57] Guha, S., Rastogi, R., and Shim, K. Cure: An efficient clustering algorithm for large databases. In *In Proceeding of the ACM SIGMOD Conference on Management of Data* (1998).

[58] Guo, B., and Nixon, M. S. Gait feature subset selection by mutual information. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 39*, 1 (2009), 36–46.

[59] Guyon, I., and Elisseeff, A. An introduction to variable and feature selection. *The Journal of Machine Learning Research 3* (2003), 1157–1182.

[60] Hailegeorgis, T. T., and Burn, D. H. Uncertainty assessment of the impacts of climate change on extreme precipitation events. Tech. rep., Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Canada, Dec. 2009. `http://www.eng.uwo.ca/research/iclr/fids/publications/cfcas-quantifying_uncertainty/reports/teklu_report.pdf`.

[61] Hajat, S., Ebi, K. L., Kovats, R. S., Menne, B., Edwards, S., and Haines, A. The human health consequences of flooding in europe: a review. In *Extreme Weather Events and Public Health Responses* (2005), Wiley-IEEE Computer Society Press, pp. 185–196.

[62] Hall, M. A., and Holme, G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering 15*, 3 (2003), 1041–4347.

[63] Hall, M. A., and Smith, L. A. Feature subset selection: A correlation based filter approach. In *Proceedings of the 4th International Conference on Neural Information Processing and Intelligent Information Systems* (Berlin, 1997), N. Kasabov, R. Kozma, K. Ko, R. O'Shea, G. Coghill, and T. Gedeon, Eds., vol. 2, Springer, pp. 855–858.

[64] Han, J., and Kamber, M. *Data Mining.* Morgan Kaufmann Publishers, 2001.

[65] Han, J., Kamber, M., and Tung, A. K. H. Spatial clustering methods in data mining: A survey. In *In Geographic Data Mining and Knowledge Discovery* (2001), Taylor and Francis.

[66] Harrald, J. R., Renda-Tanali, I., Shaw, G. L., Rubin, C. B., and Yeletaysi, S. Review of risk based priorization/decision making methodologies for dams. Tech. rep., The George Washington University, Institute for Crisis, Disaster, and Risk Management, 2004.

[67] Haykin, S. *Neural Networks: A Comprehensive Foundation. Second Edition.* Prentice Hall Inc., 1999.

[68] Haynes, H., Haynes, R. M., and Pender, G. Integrating socio-economic analysis into decision-support methodology for flood risk management at the development scale (scotland). *Water and Environment Journal 22*, 2 (2008), 117–124.

[69] Hilsenbeck, S. G., Friedrichs, W. E., Schiff, R., O'Connell, P., Hansen, R. K., Osborne, C. K., and Fuqua, S. A. W. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *Journal of the National Cancer Institute 91* (1999), 453–459.

[70] Hughes, A., Hewlett, H., Samuels, P. G., Morris, M., Sayers, P., Moffat, I., Harding, A., and Tedd, P. *Risk management for UK reservoirs, 2000.* CIRIA, London, UK, 2000.

[71] Hulme, M., Crossley, J., and Lu, X. An exploration of regional climate change scenarios for scotland. Tech. rep., Scottish Executive Central Research Unit, Edinburgh, UK, 2001.

[72] Hyvärinen, A. Survey on independent component analysis. *Neural Computing Surveys 2* (1999), 94–128.

[73] ICOLD. Dam safety guidelines - icold bulletin 59. Tech. rep., International Commission on Large Dams, 1987.

[74] IUCN. Vision for water and nature. Tech. rep., IUCN C The World Conservation Union, Cambridge, UK, 2000. `http://www.rivernet.org/general/docs/VisionWaterNature.pdf`.

[75] Jackson, J. E. *A User's Guide to Principal Components*. New York: John Wiley and Sons, 1991.

[76] Jain, A., and Flynn, P. Image segmentation using clustering. In *Advances in image understanding: A Festschrift for Azriel Rosenfeld* (1966), Wiley-IEEE Computer Society Press, pp. 65–83.

[77] Jain, A., Murty, M. N., and Flynn, P. J. Data clustering: A review. *ACM Computing Surveys 31*, 3 (1999), 264–323.

[78] Jain, A. K., and Dubes, R. C. *Algorithms for Clustering Data*. Prentice Hall Inc., New Jersey, 1988.

[79] Jebson, S. Fact sheet number 4: Climate of the united kingdom. Tech. rep., UK Meteorological Office, 2007.

[80] Jiao, A., and Yuan, L. Fault diagnosis based on adaptive genetic algorithm and bp neural network. In *2nd International Conference on Computer Engineering and Technology* (2010), Washington, DC, USA.

[81] Joachims, T. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of European Conference on Machine Learning (ECML)* (1998), vol. 1398, Springer, pp. 137–142.

[82] Johnson, B., Balseraka, P., Beaulieub, S., Cuthbertsona, B., Stewartc, R., Truesdaleb, R., Whitmoreb, R., and Younga, J. Industrial surface impoundments: environmental settings, release and exposure potential and risk characterization. *The Science of The Total Environment 317*, 1-3 (2003), 1–22.

[83] Jolliffe, I. T. *Principal Component Analysis*. Springer, 2nd edition, 2002.

[84] Jones, M. C., and Sibson, R. What is projection pursuit. *Journal of the Royal Statistical Society 150*, 1 (1987), 1–37.

[85] Kalteh, A. M., Hjorth, P., and Berndtsson, R. Review of the self-organizing map (som) approach in water resources: analysis, modeling and application. *Environmental Modelling & Software 23*, 7 (2008), 835–845.

[86] Karamouz, M., Kerachian, R., Akhbari, M., and Hafez, B. Design of river water quality monitoring networks: a case study. *Environmental Modeling and Assessment 14*, 4 (2009), 705–714.

[87] Kay, A. L., Jones, R. G., and Reynard, N. S. Rcm rainfall for uk flood frequency estimation. ii. climate change results. *Journal of Hydrology 318*, 1-4 (2006), 163–172.

[88] Keef, C., Tawn, J., and Svensson, C. Spatial risk assessment for extreme river flows. *Journal of the Royal Statistical Society*.

[89] Kelman, I. Defining risk. *FloodRiskNet Newsletter 2* (2003), 6–8.

[90] Kidson, R., and Richards, K. S. Flood frequency analysis: assumptions and alternatives. *Progress in Physical Geography 29*, 3 (2005), 392–410.

[91] Kira, K., and Rendell, L. A. A practical approach to feature selection. In *Proceedings 9th International Conference on Machine Learning* (San Francisco, CA, USA, 1992), N. Kasabov, R. Kozma, K. Ko, R. O'Shea, G. Coghill, and T. Gedeon, Eds., Morgan Kaufmann Publishers Inc., pp. 249–256.

[92] Kohavi, R., and John, G. H. Wrappers for feature subset selection. *Artificial Intelligence 97*, 1-2 (1997), 273–324.

[93] Kohonen, T. The self-organizing map. *proceedings of the IEEE 78*, 9 (1990), 1464–1480.

[94] Kohonen, T. *Self-organizing Maps*, 3rd ed. Springer, Berlin, Germany, 2001.

[95] Kononenko, I. Estimating attributes: Analysis and extensions of relief. In *Proceedings of the Seventh European Conference on Machine Learning* (Berlin, 1994), F. Bergadano, Ed., vol. 784, Springer, pp. 171–182.

[96] Kuhn, G., and Diekmann, B. Data report: Bulk sediment composition, grain-size, clay, and silt mineralogy of pleistocene sediments from odp leg 177 sites 1089 and 1090. *Proc. ODP, Sci. Results 177* (2003), 1–10.

[97] Labib, K., and Vemuri, V. R. An application of principal component analysis to the detection and visualization of computer network attacks. *Annales of Telecommunications 61* (2006), 225–234.

[98] Larinier, M. Dams and fish migration. *FAO Fisheries Technical Paper 419* (2000), 45–89.

[99] Lee, B.-H., and Scholz, M. Application of the self-organizing map (som) to assess the heavy metal removal performance in experimental constructed wetlands. *Water Research 40*, 18 (2006), 3367–3374.

[100] Lehner, B., and Döll, P. The human health consequences of flooding in europe: a review. In *EuroWasser - Model-based assessment of European water resources and hydrology in the face of global change* (2001), University of Kassel, pp. 6.1–6.14.

[101] Lewis M. Cowardin, Virginia Carter, F. C. G., and LaRoe, E. T. Classification of wetlands and deepwater habitats of the united states. Tech. rep., U.S. Department of the Interior, Fish and Wildlife Service, Washington, D.C. Jamestown, ND. Northern Prairie Wildlife Research Center Online.

[102] Li, T., and Ogihara, M. Toward intelligent music information retrieval. *IEEE Transactions on Multimedia 8*, 3 (2006), 564–574.

[103] Liu, H., and Setiono, R. A probabilistic approach to feature selection - a filter solution. In *Proceedings of the 13th International Conference on Machine Learning* (1996), F. Bergadano, Ed., vol. 784, Morgan Kaufmann, pp. 319–327.

[104] Liu, H., and Yu, L. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering 17*, 3 (2005), 491–502.

[105] Liu, X., Wu, J., and Xu, J. Characterizing the risk assessment of heavy metals and sampling uncertainty analysis in paddy field by geostatistics and gis. *Environmental Pollution 141*, 2 (2006), 257–264.

[106] Liu, Y., and Weisberg, R. Patterns of ocean current variability on the west florida shelf using the self-organizing map. *Journal of Geophysical Research 110*, C06003 (2005).

[107] Liu, Y., and Weisberg, R. H. A review of self-organizing map applications in meteorology and oceanography. In *Computer Music Modeling and Retrieval-Lecture Notes in Computer Science* (2001), InTech, pp. 253–272.

[108] Liu, Y., Weisberg, R. H., and Mooers, C. N. K. Performance evaluation of the self-organizing map for feature extraction. *Journal of Geophysical Research 111*, C05018 (2006).

[109] Lochhead, R., and Cunningham, R. Reservoir safety in scotland-a consultation document. Tech. rep., Scottish Government, Edinburgh, UK, Jan. 2010. `http://www.scotland.gov.uk/Resource/Doc/299725/0093437.pdf`.

[110] Long, R., Vawter, N., Horn, C., Aronson, W., Connell, P., Hill, J., Fugo, A., Gardner, J., Thornton, A., Jungwirth, M., Powell, S., and Kelly, C. *The Hydrologic Water Cycle - The Water Sourcebook Series 9-12*. Legacy, Inc., 1994.

[111] Loures, L., and Panagopoulos, T. From derelict industrial areas towards multifunctional landscapes and urban renaissance. *Wseas Transactions on Environment and Development 3*, 10 (2007), 181–188.

[112] MacQueen, J. B. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (1967), vol. 1, University of California Press, pp. 281–297.

[113] Mardia, K. V., Kent, J. T., and Bibby, J. M. *Multivariate Analysis (Probability and Mathematical Statistics)*. Academic Press, 1995.

[114] McCallum, A. K. Multi-label text classification with a mixture model trained by em. In *Proceedings of the AAAI' 99 Workshop on Text Learning* (1999), Orlando, Florida, USA, pp. 1–7.

[115] McCulloch, W. S., and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5* (1943), 115–133.

[116] McLemore, V. T. Chicosa lake state park. *New Mexico Geology 10*, 3 (1988), 62–64.

[117] Mihael Ankerst, Markus M. Breunig, H.-P. K. J. S. Optics: Ordering points to identify the clustering structure. In *In Proceeding of ACM SIGMOD99 Internatioanl Conference on Management of Data* (1999).

[118] Millerick, A. Validation of fsr and feh depth/duration/frequency using recent met office rainfall data. In *Proceeding of the 3rd CIWEM National Conference* (Wakefield, West Yorkshire, UK, 2005). `http://www.microdrainage.co.uk/assets/documents/ValidationFSRFEH.pdf`.

[119] Mitchell, G. N. Water quality issues in the british uplands. *Applied Geography 11*, 3 (1991), 201–214.

[120] Mohssen, M. An insight into flood frequency for design floods. In *1st International Conference on Flood Recovery, Innovation and Response (FRIAR)* (2008), 155–164.

[121] Montaldo, N., Mancini, M., and Rosso, R. Flood hydrograph attenuation induced by a reservoir system: analysis with a distributed rainfall-runoff model. *Hydrological Processes 18*, 3 (2004), 543–563.

[122] Moss, B. *Ecology of Fresh Waters: Man and Medium*. Blackwell Science, Oxford, UK, 1988.

[123] Nayak, R., Jain, L. C., and Ting, B. K. H. Artificial neural networks in biomedical engineering: a review. In *In Proceedings Asia-Pacific Conference on Advance Computation* (2001), Sidney, Australia, pp. 887–892.

[124] Nisbet, T. R., Welch, D., and Doughty, R. The role of forest management in controlling diffuse pollution from the afforestation and clearfelling of two public water supply catchments in argyll, west scotland. *Forest Ecology and Management 158*, 1-3 (2002), 141–154.

[125] Oliver, M. A., and Webster, R. How geostatistics can help you. *Soil Use and Management 7*, 4 (1991), 206–217.

[126] Oliver, M. A., Webster, R., and McGrath, S. P. Disjunctive kriging for environmental management. *Environmetrics 7*, 3 (1996), 333–357.

[127] Parlon, T., Benton, S., Smyth, T., andDonal Buckley, J. B., and andJohn Tiernan, M. T. Report of the flood policy review group. Tech. rep., Office of Publick Works, 2003.

[128] Parry, M. L., Canziani, O. F., Palutikof, J. P., van der Linden, P. J., and Hanson, C. E. *Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC)*. Cambridge University Press, Cambridge, UK, 2007.

[129] Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine 2*, 6 (1945), 559–572.

[130] Peng, H., and Long, F. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 8 (2005), 1226–1238.

[131] Pitt, S. M. Learning lessons from the 2007 floods - the pitt review. Tech. rep., Cabinett Office, 2008.

[132] POST. River basin management plans. Tech. rep., Parliamentary Office of Science and Technology, London, UK, 2008. `http://www.parliament.uk/documents/post/postpn320.pdf`.

[133] Pratap, R. *Getting Started with MATLAB: A Quick Introduction for Scientists and Engineers*. Oxford University Press, Oxford, 2002.

[134] Qi, G.-J., Hua, X.-S., Rui, Y., Tang, J., Mei, T., and Zhang, H.-J. Correlative multi-label video annotation. In *Proceedings of the 15th international conference on Multimedia* (2007), New York, NY, USA: ACM Press, pp. 17–26.

[135] Quinlan, J. R. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Mateo, CA, 1993.

[136] Ramchunder, S. J., Brown, L. E., and Holden, J. Environmental effects of drainage, drain-blocking and prescribed vegetation burning in uk upland peatlands. *Progress in Physical Geography 33*, 1 (2009), 49–79.

[137] Ramos-Scharroon, C. E., and MacDonald, L. H. Measurement and prediction of natural and anthropogenic sediment sources, st. john, us virgin islands. *Catena 71*, 2 (2007), 250–266.

[138] Raychaudhuri, S., Stuart, J. M., and Altman, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing 5* (2000), 452–463.

[139] RCS. The ramsar convention manual: a guide to the convention on wetlands (ramsar, iran, 1971), 4th edition. Tech. rep., Ramsar Convention Secretariat.

[140] Rettemeier, K., and Köngeter, J. Dam safety management: Overview of the state of the art in germany compared to other european countries. In *Proceedings of the International Symposium on New Trends and Guidelines on Dam Safety* (1998), vol. 1, pp. 55–62.

[141] Richard O. Duda, Peter E. Hart, D. G. S. *Pattern Classification. Second Edition*. John Wiley & Sons Inc., 2001.

[142] Rifkin, R., and Klautau, A. In defence of one-versus-all classification. *Journal of Machine Learning Research 5* (2004), 101–141.

[143] Rish, I. An empirical study of the naïve bayes classifier. In *Proceedings of IJCAI-01 workshop on Empirical Methods in AI, International Joint Conference on Artificial Intelligence* (2001), pp. 41–46.

[144] Rissler, P. Dimensioning of the design flood as part of a reservoir safety concept. *International Journal on Hydropower and Dams 8* (2001), 98–107.

[145] Rivoirard, J. *Introduction to Disjunctive Kriging and Non-linear Geostatistics.* Clarendon Press, Oxford, England, United Kingdom, 1994.

[146] Rubin, A. P. D. N. M. L. D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*, 1 (1977), 1–38.

[147] Rui, Y., Huang, T. S., and Chang, S.-F. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation 10*, 4 (1999), 39–62.

[148] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning internal representations by error propagation. In *Parallel Distributed Processing of Explorations in the Microstructure of Cognition* (1986), Cambridge, MA: MIT Press, pp. 318–362.

[149] Rustum, R., Adeloye, A. J., and Scholz, M. Applying kohonen self-organizing map as a software sensor to predict biochemical oxygen demand. *Water Environment Research 80*, 1 (2008), 32–40.

[150] Saaty, T. L. Deriving the ahp 1-9 scale from first principles. In *In Proceedings of ISAHP 2001* (2001), Bern, Switzerland.

[151] Saeys, Y., Inza, I., and Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics 23*, 19 (2007), 2507–2517.

[152] Salton, G. Developments in automatic text retrieval. *IEEE Transactions on Multimedia 253*, 5023 (1991), 974–980.

[153] Sanderson, D. Cities, disasters and livelihoods. *Environment and Urbanization 12*, 2 (2000), 93–102.

[154] Sauquet, E. Mapping mean annual river discharges: geostatistical developments for incorporating river network dependencies. *Journal of Hydrology 331*, 1-2 (2006), 300–314.

[155] SAWA. Strategic alliance for water management actions (sawa) project information. `http://www.sawa-project.eu/`, July 2011.

[156] Sayers, P., Hall, J., and Meadowcroft, I. Towards risk-based flood hazard management in the u.k. In *Proceedings of the Institution of Civil Engineers* (2002), vol. 150, Thomas Telford Ltd., pp. 36–42.

[157] Schapire, R. E., and Singer, Y. Boostexter: a boosting-based system for text categorization. *Machine Learning 39*, 2/3 (2000), 135–168.

[158] Scholz, M. *Wetland Systems to Control Urban Runoff.* Elsevier, Amsterdam, The Netherlands, 2006.

[159] Scholz, M. Classification methodology for sustainable flood retention basins. *Landscape and Urban Planning 81*, 3 (2007), 246–256.

[160] Scholz, M. Expert system outline for the classification of sustainable flood retention basins (sfrbs). *Civil Engineering and Environmental Systems 24*, 3 (2007), 193–209.

[161] Scholz, M. Classification of flood retention basins: The kaiserstuhl case study. *Environmental and Engineering Geoscience 24*, 2 (2008), 61–80.

[162] Scholz, M. *Wetland Systems - Storm Water Management Control*. Springer Verlag, Berlin, Germany, 2010.

[163] Scholz, M., and Sadowski, A. J. Conceptual classification model for sustainable flood retention basins. *Journal of Environmental Management 90*, 1 (2009), 624–633.

[164] Schumacher, B. A. Methods for the determination of total organic carbon (toc) in soils and sediments. Tech. rep., Ecological Risk Assessment Support Center, Office of Research and Development, US Environmental Protection Agency, 2002. `http://www.epa.gov/esd/cmb/research/papers/bs116.pdf`.

[165] Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys 34*, 1 (2002), 1–47.

[166] SEPA. Indicative river and coastal flood map (scotland)-summary of technical methodology. Tech. rep., Scottish Environmental Protection Agency, Stirling, UK, Sep. 2006. `http://www.scotland.gov.uk/Resource/Doc/299725/0093437.pdf`.

[167] SEPA. Scotland's wfd aquatic monitoring strategy. Tech. rep., The Scottish Environment Protection Agency (SEPA), 2007.

[168] Shah, H., Undercoffer, J., and Joshi, A. Fuzzy clustering for intrusion detection. In *Proceeding of the 12th IEEE International Conference on Fuzzy Systems* (2003), vol. 2.

[169] Shepherd, T., Chenery, S., Pashley, V., Lord, R. A., Ander, L., Breward, N., Hobbs, S., and Horstwood, M. Regional lead isotope study of a polluted river catchment: River wear, northern england, uk. *Science of the Total Environment 407*, 17 (2009), 4882–4893.

[170] Snoek, C. G., Worring, M., van Gemert, J. C., Geusebroek, J.-M., and Smeulders, A. W. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th annual ACM international conference on Multimedia* (2006), New York, NY, USA: ACM Press, pp. 421–430.

[171] Spachinger, K., Dorner, W., Fuchs, S., Serrhini, K., and Metzka, R. Flood risk and flood hazard maps - visualisation of hydrological risks. In *XXIVth Conference of the Danubian Countries, Institute of Physics (IOP) Conference*

Series: Earth and Environmental Science (2008), vol. 4, IOP Publishing Ltd, pp. 1–17.

[172] Staniford-Chen, S., and Heberlein, L. Holding intruders accountable on the internet. In *Proceedings of the 1995 IEEE Symposium on Security and Privacy* (1995), Washington, DC, USA, IEEE Computer Society, pp. 39–49.

[173] Stevens, C. J., and Quinton, J. N. Diffuse pollution swapping in arable agricultural systems. *Critical Reviews in Environmental Science and Technology 39*, 6 (2009), 478–520.

[174] Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B (Methodological) 36*, 2 (1974), 111–147.

[175] Sutherland, W. J. *Ecological Census Techniques - a Handbook*, 2rd edition ed.

[176] Taylor, C., and Alves-Foss, J. Nate: Network analysis of anomalous traffic events, a low-cost approach. In *In Proceedings of the New Security Paradigms Workshop 2001* (2002), Cloudcroft, New Mexico, USA, pp. 89–96.

[177] Timlett, R., and Gordon-Walker, S. Dealing with the deluge - urban water management in a changing climate. Tech. rep., WWF and RSA, 2010. `http://www.wwfrsapartners.com/static/uploads/page_files/WWFRSA_SuDsReportFINAL.pdf`.

[178] Tipping, M. E., and Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B 61*, 3 (1999), 611–622.

[179] Trohidis, K., Tsoumakas, G., Kalliris, G., and Vlahavas, I. Multilabel classification of music into emotions. In *Proceedings of 9th International Conference on Music Information Retrieval (ISMIR 2008)* (2008), Philadelphia, PA, USA, pp. 325–330.

[180] Tsoumakas, G., and Katakis, I. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining 3*, 3 (2007), 1–13.

[181] Tsoumakas, G., Katakis, I., and Vlahavas, I. Multilabel classification of music into emotions. In *in: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook, 2nd edition* (2010), Springer, pp. 667–685.

[182] Ukonmaanaho, L., Nieminen, T., Rausch, N., Cheburkin, A., Roux, G. L., and Shotyk, W. Recent organic matter accumulation in relation to some climatic factors in ombrotrophic peat bogs near heavy metal emission sources in finland. *Global and Planetary Change 53*, 4 (2006), 259–268.

[183] van Duivendijk, J. Assessment of flood management options. Tech. rep., World Commission on Dams, 1999). `http://oldwww.wii.gov.in/eianew/eia/dams%20and%20development/kbase/contrib/opt173.pdf`.

[184] Vembu, S., and Baumann, S. A self-organizing map based knowledge discovery for music recommendation systems. In *Computer Music Modeling and Retrieval-Lecture Notes in Computer Science* (2005), vol. 3310, pp. 119–129.

[185] Vesanto, J., Himberg, J., Alhoniemi, E., and Parhankangas, J. Som toolbox for matlab 5 documentation. Tech. rep., Helsinki University of Technology, Helsinki, Finland, 2000. `http://www.cis.hut.fi/projects/somtoolbox/`.

[186] Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association 58*, 301 (1963), 236–244.

[187] Watzin, M. C., and McIntosh, A. W. Aquatic ecosystems in agricultural landscapes: a review of ecological indicators and achievable ecological outcomes. *Journal of Soil and Water Conservation 54*, 4 (1999), 636–644.

[188] Webster, R., and Oliver, M. A. *Geostatistics for Environmental Scientists*. John Wiley and Sons, Chichester, England, United Kingdom, 2001.

[189] Werritty, A., Black, A., Duck, R., Finlinson, B., Thurston, N., Shackley, S., and Crichton, D. Climate change: flooding occurrences review. Tech. rep., Scottish Executive Central Research Unit, Edinburgh, UK, 2002. `http://www.scotland.gov.uk/Resource/Doc/156664/0042098.pdf`.

[190] Woltjer, J., and Kranen, F. Articulating resilience in flood risk management and spatial planning. In *24th AESOP Annual Conference* (2010), no. III-23, Helsinki, Finland, pp. 1–25.

[191] Woods-Ballard, B., Kellagher, R., Martin, P., Jefferles, C., Bray, R., and Shaffer, P. The suds manual (c697). Tech. rep., CIRIA, London, UK, 2007. `http://www.ceh.ac.uk/Feh2/InsidetheFloodEstimationHandbook.html`.

[192] Yevjevich, V. Technology for coping with floods in the 21st century. In *Coping with Floods, NATO ASI Series E: Applied Sciences* (1994), vol. 257, Washington, DC, USA, pp. 36–44.

[193] Zhang, M.-L., and Zhou, Z.-H. Multi-label neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering 18*, 10 (2006), 1338–1351.

[194] Zhang, M.-L., and Zhou, Z.-H. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recogition 40*, 7 (2007), 2038–2048.

[195] Zhang, Y., Burer, S., and Street, W. N. Ensemble pruning via semi-definite programming. *Journal of Machine Learning Research 7* (2006), 1315–1338.

[196] Zhang, Y., van Dijk, M. A., Liu, M., Zhu, G., and Qin, B. The contribution of phytoplankton degradation to chromophoric dissolved organic matter (cdom) in eutrophic shallow lakes: field and experimental evidence. *Journal of Soil and Water Conservation 43*, 18 (2009), 4685–4697.

[197] Zhu, L., Yang, J., and Shen, H.-B. Multi label learning for prediction of human protein subcellular localizations. *The Proteon Journal 28*, 9-10 (2009), 384–390.

Appendix A: Detailed Description of Variables with corresponding boundary conditions

Site reference number:	Full site name:	Template version: 041108
Pictures taken (yes or no):	Picture numbers:	Date:

EU INTERREG SAWA PROJECT on Flood Retention Structures: Classification variables for Sustainable Flood Retention Basins (SFRB) to control runoff and diffuse pollution in Scotland

Variable (unit)	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5
1. Engineered (%)	Mature, permanently filled and man-made structure (e.g. dam) with full engineering control; i.e. similar to a traditional drinking water reservoir (>50)	Mostly permanently filled and man-made traditional structure (>45 to 50) with some natural elements, but virtually full engineering control	Predominantly engineered structure (25 to 45) fitted well into the natural landscape with predominantly passive control (e.g. simple outlet)	Aesthetically pleasing natural and occasionally dry formation with some engineered (20 to <25) features (e.g. outlet), but relatively natural channel base	Almost entirely natural and occasionally dry formation including a wide natural channel base (<20 engineered; e.g. inlet and/or outlet)
Value (%):					
Confidence (%):					
Comment:					
2. Dam Height (m)	Relatively high dam (measured from bottom of water body to top of dam) as the dominant structure (>8)	Part of engineered structure featuring a dam (>5.5 to 8)	Small dam and natural containment combined (3.5 to 5.5)	Ecologically engineered and aesthetically pleasing structure with a small dam (1.5 to <3.5)	Insignificant dam or no dam but natural containment of water due to topography; e.g. in a valley (<1.5)
Value (m):					
Confidence (%):					
Comment:					
3. Dam Length (m)	Very long dam (>900)	Long dam (>700 to 900)	Normal dam length (100 to 700)	Short dam (60 to <100)	Very short dam or no artificial dam(<60)
Value (m):					
Confidence (%):					
Comment:					
4. Outlet Arrangement and Operation (%)	Most likely combined outlets; engineered spillway; most likely fully-automatic operation (>85)	Combined or separate outlets; engineered spillway; fully-automatic operation likely (>75 to 85)	Combined or separate outlets; potentially a spillway present; partially-automatic operation likely (15 to 75)	Separate outlets; at least one outflow is fixed; possibly manual operation (8 to <15)	Separate outlets (if any); fixed base flow and/or flood flow outlets; probably no engineered operation mechanism (<8)
Value (%):					
Confidence (%):					
Comment:					
5. Aquatic Animal Passage (%)	Unhindered aquatic animal passage; e.g. a fish ladder is present (>80)	Easy aquatic animal passage; e.g. only a short below-ground outlet causes a minor obstruction, and an old fish ladder might be present (>60 to 80)	Standard aquatic animal passage; some pipes and relative wide dam (20 to 60)	Problematic aquatic animal passage; long inlet and/or outlet pipes; isolated habitat (10 to <20)	Very problematic aquatic animal passage; e.g. very long inlet and outlet pipes; potentially a waterfall present; isolated and hostile habitat (<10)
Value (%):					
Confidence (%):					
Comment:					

6. Land Animal Passage (%)	Unhindered land animal passage; virtually no physical obstractions (>70)	Easy land animal passage; e.g. only a small dam causes a minor obstruction (>50 to 70)	Standard land animal passage (20 to 50)	Problematic land animal passage; isolated habitat and high physical structures (10 to <20)	Very problematic land animal passage; e.g. isolated and hostile habitat such as a town, and high concrete dam (<10)
Value (%):					
Confidence (%):					
Comment:					
7. Floodplain Elevation (m)	Retention basin predominantly elevated; e.g. valley in the highlands with normal slopes (>2)	Basin elevated and not well integrated into the landscape (>1.25 to 2)	Basin elevated but well-integrated into the landscape (1 to 1.25)	Retention basin partly elevated and structure perfectly integrated into the landscape (0.5 to <1)	Natural retention basin not elevated and virtually level with the inflow source (e.g. stream); most drinking water reservoirs (<0.5)
Value (m):					
Confidence (%):					
Comment:					
8. Basin & Channel Connectivity (m)	Very long connectivity (i.e. basin and watercourse are far away (>20) from each other)	Long connectivity (>15 to 20)	Short connectivity (5 to 15), but elements are clearly separated from each other	Short connectivity; i.e. offline (1 to <5); well-integrated into the landscape	Basin directly connected (<1) with the watercourse (i.e. basin and stream are virtually online)
Value (m):					
Confidence (%):					
Comment:					
9. Wetness (%)	Very wet basin (almost entirely submerged; i.e. >45); similar to a pond or drinking water supply reservoir	Wet basin with minor natural components such as reeds near embankments (>35 to 45)	Partly dry and partly wet basin with minor natural components (25 to 35)	Predominantly dry basin or pond, but with natural components (5 to <25)	Very dry area with natural components; i.e. dry basin or pond (<5)
Value (%):					
Confidence (%):					
Comment:					
10. Proportion of Flow within Channel (%)	Permanently flooded main channel through basin; i.e. online basin with no engineered bypass (>99)	Permanently flooded main channel through basin (>95 to 99) and very occasionally flooded bypass	Partly flooded main channel (90 to 95) and occasionally partly flooded bypass	Main channel (80 to <90) and partly flooded bypass	Main channel (<80) and bypass (virtually offline basin) taking most of the flood water compared to the water source
Value (%):					
Confidence (%):					
Comment:					
11. Mean Flooding Depth (m)	Mostly very deep flooding depth (including normal water depth) as in reservoirs (>4); virtually permanently wet	Deep flooding depth (>2 to 4)	Partly flooded; normal flooding depth (0.9 to 2)	Partly flooded; shallow flooding depth (0.4 to <0.9)	Only occasionally and partly flooded; very shallow flooding depth (<0.4); virtually a dry basin
Value (m):					
Confidence (%):					
Comment:					

12. Typical Wetness Duration (d a^{-1})	Very long flooding event due to intense and long storms (>350); potentially permanently flooded (e.g. pond or reservoir)	Prolonged flooding event due to long storms (>200 d to 350)	Occasional flooding event (20 to 200)	Mostly short flooding event (5 d to <20)	Mostly very short flooding events (<5); virtually a permanently dry basin or pond
Value (d a^{-1}):					
Confidence (%):					
Comment:					
13. Estimated Flood Frequency (a^{-1})	Very high flood frequency due to high storm frequencies (>12) leads to serious build environment destruction	High flood frequency due to high storm frequencies (>9 to 12) leads to landscape damage	Normal flood frequency (5 to 9) controlled by management strategies and structures	Low flood frequency, which does not lead to any un-aesthetical sights (1 to <5)	Very low flood frequency (<1), rarely notices, monitored and recorded
Value (y^{-1}):					
Confidence (%):					
Comment:					
14. Basin Bed Gradient (%)	Very high gradient (usually measured from inlet to outlet); possibly also steep valley slopes (>13)	High gradient (>8 to 13)	Normal gradient between inlet and outlet (4 to 8)	Low gradient (2.5 to <4)	Very low gradient; flat valley slopes or lowland area (<2.5)
Value (%):					
Confidence (%):					
Comment:					
15. Mean Basin Flood Velocity (cm s^{-1})	Mostly very high mean velocity due to intense storms and compact structure; channel bed dominates the basin (>150)	High velocity; very few structural elements that slow down water (>125 to 150)	Variable velocity within channel and basin due to complex structure (65 to 125)	Low velocity due to complex structure(45 to <65)	Mostly very low velocity but rarely stagnant; basin is complex and/or large in comparison to the channel (<45)
Value (cm s^{-1}):					
Confidence (%):					
Comment:					
16. Wetted Perimeter (m)	Very long wetted perimeter (>1100); potentially a very variable natural structure including islands and vegetation	Aesthetically pleasing long wetted perimeter (>850 to 1100)	Normal wetted perimeter length; possibly identical to the embankment (200 to 850)	Short wetted perimeter (90 to <200)	Very short wetted perimeter; potentially a man-made simple structure (<90)
Value (m):					
Confidence (%):					
Comment:					
17. Maximum Flood Water Volume (m^3)	Very high volume; basin was designed for a very rare storm event; depth includes sediment and base water level (>500,000)	High volume (>100,000 to 500,000)	Normal maximum flood water volume (50,000 to 100,000)	Low volume (15,000 to <50,000)	Very low volume; depth includes sediment and base water level (<15,000)
Value (m^3):					
Confidence (%):					
Comment:					

18. Flood Water Surface Area (m^2)	Very large surface area reflecting a very large basin; includes parts of the flooded floodbanks (>100,000)	Large surface area reflecting a large catchment (>60,000 to 100,000)	Normal surface area covered with water (2,000 to 60,000)	Small aesthetically pleasing surface area (400 to <2,000)	Very small surface area reflecting a very small catchment area (<400)
Value (m^2):					
Confidence (%):					
Comment:					
19. Mean Annual Rainfall (mm)	Very high precipitation due to intense and long storms in the catchment; highland climate (>1000)	High precipitation due to long storms in the catchment; partly highland climate (>900 to 1000)	Normal mean annual precipitation in the catchment; wet temperate climate (800 to 900)	Relatively low precipitation in the catchment; temperate lowland climate; e.g. Edinburgh (750 to >800)	Very low precipitation in the catchment; sometimes semi-arid climate (<750)
Value (mm):					
Confidence (%):					
Comment:					
20. Drainage (cm d^{-1})	Very well-drained with a very high estimated infiltration rate (>10)	Well-drained with a relatively high infiltration rate (>5 to 10)	Drained with a moderate infiltration rate (1 to 5)	Not well-drained with a low estimated infiltration rate (0.5 to <1)	Mostly not drained (<0.5); usually a boggy area
Value (mm):					
Confidence (%):					
Comment:					
21. Impermeable Soil Proportion (%)	Very high proportion of rock, clay and/or other impermeable material present (>40)	High proportion of clay and/or rock (>25 to 40)	Variable proportions of impermeable soil (e.g., clay) present (5 to 25)	Low proportion of impermeable material such as clay present (2 to <5)	Clay and other impermeable material virtually absent (<2)
Value (%):					
Confidence (%):					
Comment:					
22. Seasonal Influence (%)	Strong seasonal influence; e.g., typical temperate climate in mountains (>80)	Most seasonal influence; e.g., area around Freiburg in Baden, Germany (>60 to 80)	Moderate seasonal influence; e.g., area around Edinburgh, Scotland (55 to 60)	Weak seasonal influence (45 to <55)	Very weak seasonal influence; e.g., some Mediterranean countries(<45)
Value (%)					
Confidence (%):					
Comment:					
23. Site Elevation (m)	Very high elevation (basin bottom is the reference point); e.g., highlands (>400)	High elevation (>350 to 400)	Typical elevation; no highlands and no lowlands (280 to 350)	Low elevation (150 to <280)	Very low elevation; e.g., marshes (<150)
Value (m):					
Confidence (%):					
Comment:					
24. Vegetation Cover (%)	Vegetated basin (not catchment) area with high roughness (e.g., strong reed stands and/or mature trees such as willows); >80	Vegetated area with moderate roughness (e.g., predominantly reeds, bushes and some trees); >60 to 80	Vegetated area with low roughness (e.g., bushes and some reeds); early succession of plants; 30 to 60	Partly not vegetated area (e.g., some grassland or submerged aquatic vegetation); 10 to <30	Predominantly not vegetated basin (not catchment); e.g., only short-mowed lawn or tarmac cover; <10
Value (%)					
Confidence (%):					
Comment:					

25. Algal Cover in Summer (%)	Very dominant algal bloom likely(>70)	Dominant algal bloom likely(>50 to 70)	Normal algal presence likely (10 to 50)	Small algal presence likely (3 to <10)	Virtually no algae present; blooms are very unlikely (<3)
Value (%):					
Confidence (%):					
Comment:					
26. Relative Total Pollution (%)	Virtually permanently polluted (high organic and inorganic solids content; occasionally sewage), but still acceptable for some economic use (>60)	Occasionally polluted with solids and/or organics during runoff events (<35 to 60)	Occasionally partly polluted during some runoff events; sometimes good water quality (15 to 35)	Minor occasional pollution; good water quality; environment may still be rich in species (7 to <15)	Rarely minor pollution; very good water quality; environment is rich in species (<7)
Value (%):					
Confidence (%):					
Comment:					
27. Mean Sediment Depth (cm)	Very deep and potentially natural sediment layer (>9) present (mature and stable system); most likely organic	Deep sediment layer (>6 to 9) present (mature system)	Still an aesthetically pleasing sediment layer (2 to 6); occasional basin management	Occasional presence of sediment (0.5 to <2) due to passive and/or occasional basin management	Occasional presence of sediment (<0.5) due to passive and/or occasional basin management
Value (cm):					
Confidence (%):					
Comment:					
28. Organic Sediment Proportion (%)	Relatively deep organic and potentially natural sediment layer; resulting from organic decay over decades (>80)	Organic and predominantly man-made sediment layer (>60 to 80)	Mixture between inorganic and organic sediment layer (>40 to 60 organic)	Predominantly inorganic sediment layer (20 to 40 organic), which is sometimes managed/removed regularly	Relatively deep inorganic sediment (e.g. river sand) layer (<20 organic) requiring regular removal
Value (%)					
Confidence (%):					
Comment:					
29. Flotsam Cover (%)	Serious flotsam cover causing hydraulic problems; immediate removal required (>80)	Troublesome flotsam cover; regular removal required (>70 to 80)	Significant presence of flotsam; some minor flow obstruction (30 to 70)	Occasional presence of flotsam; very minor flow obstruction (10 to <30)	Virtually no or no flotsam (e.g. branches, weeds and rubbish); no flow obstruction (<10)
Value (%)					
Confidence (%):					
Comment:					
30. Catchment Size (km^2)	Very large catchment; partly difficult to define; some major tributaries possible (>30)	Large catchment; tributaries possible (>15 to 30)	Typical catchment size (2 to 15)	Small catchment; small inflow sources (0.5 to <2)	Very small catchment; no clear inflows sources; difficult to identify on a map (<0.5)
Value (km^2):					
Confidence (%):					
Comment:					

31. Urban Catchment Proportion (%)	Very high urban catchment proportion (>65); e.g., city, town, rail and motorways	High urban catchment proportion (>45 to 65); e.g., town (or large villages) and major roads	Significant urban catchment proportion (15 to 45); e.g., large villages and some major roads	Insignificant urban catchment proportion (4 to <15); e.g., small villages and minor roads	Not urbanized catchment (<4); e.g., only individual houses and/or farms
Value (%):					
Confidence (%):					
Comment:					
32. Arable Catchment Proportion (%)	Very highly intensively used (mostly crops in very large fields) agricultural catchment proportion (>50)	Highly intensive used (mostly crops; some fruit trees) agricultural catchment proportion (>35 to 50)	Significant arable catchment proportion as can be expected, e.g. in central Europe (8 to 35)	Insignificant agricultural catchment proportion; predominantly arable fields and gardens (4 to <8)	Not agriculturally used; some isolated arable fields and gardens with bare soil (<4)
Value (%):					
Confidence (%):					
Comment:					
33. Pasture Catchment Proportion (%)	Very highly intensively used pasture catchment proportion; plenty of cattle and related animals visible (>85)	Highly intensively pasture catchment proportion; cows and sheep visible (>75 to 85)	Significant pasture catchment proportion (30 to 75%); typical for, e.g., central Europe	Insignificant agricultural catchment proportion; predominantly pastures used for grazing (10 to <30)	Not agriculturally used; some isolated pastures used for grazing or landscape purposes (<10)
Value (%):					
Confidence (%):					
Comment:					
34. Viniculture Catchment Proportion (%)	Very highly intensively used viniculture catchment proportion (>85)	Highly intensively viniculture catchment proportion (>75 to 85)	Significant viniculture catchment proportion (30 to 75%); typical for viniculture areas	Almost insignificant viniculture catchment proportion (10 to <30)	Not agriculturally used; none or some isolated vineyards not used for commercial purposes (<10)
Value (%):					
Confidence (%):					
Comment:					
35. Forest Catchment Proportion (%)	Very high forested proportion outside the built environment(>40)	High forested proportion (>20 to 40)	Significant forested proportion (7 to 20)	Virtually an insignificant forested catchment proportion; some trees in a partly built environment (3 to <7)	Not used for forestry purposes but some minor vegetation in, e.g. built up areas (<3)
Value (%):					
Confidence (%):					
Comment:					
36. Natural Catchment Proportion (%)	Very high natural catchment proportion outside the built environment; e.g. heather in the uplands (>30)	High natural catchment proportion; e.g. heather in the uplands (>15 to 30)	Significant natural catchment proportion (7 to 15)	Virtually an insignificant natural catchment proportion; some semi-natural vegetation in a partly built environment (2 to <7)	Some minor vegetation in, e.g. built up areas (<2)
Value (%):					
Confidence (%):					
Comment:					

37. Groundwater Infiltration (%)	Virtually the entire retention basin depends on groundwater infiltration to be wet (>50)	Very high groundwater infiltration (>40 to 50)	High groundwater infiltration (10 to 40)	Measurable groundwater infiltration from, e.g., springs (5 to <10)	Not significantly groundwater-fed (<5), but may exfiltrate into the local aquifer
Value (%):					
Confidence (%):					
Comment:					
38. Mean Depth of the Basin (m)	Very deep; most likely to be a large natural water body (>50)	Deep; likely to be a natural water body or a reservoir (>20 to 50)	Significant depth; not defined as a wetland anymore; could be a reservoir (7 to 20)	Shallow; could be a wetland if submerged (1 to <7)	Very shallow; likely to be a planted wetland if submerged (<1)
Value (m):					
Confidence (%):					
Comment:					
39. Length of Basin (m)	Very long; most likely a natural water body such as a lake (>2000)	Long; likely to be a lake (>1000 to 2000)	Significant length (500 to 1000)	Short (100 to <500)	Very short water body (<100)
Value (m):					
Confidence (%):					
Comment:					
40. Width of Basin (m)	Very wide; most likely a natural water body such as a lake (>1000)	Wide; likely to be a lake (>500 to 1000)	Significant width (200 to 500)	Narrow (50 to <200)	Very narrow water body (<50)
Value (m):					
Confidence (%):					
Comment:					

Bias and purpose (values should add up to 100):

Dominant **hydraulic purposes** (i.e. water storage reservoir; flood protection; efficient hydraulic water management)	**Drinking water** supply (typical potable water storage reservoir)	**Production industry** water supply (e.g., mills or water for the chemical industry)	**Sustainable drainage** (i.e. best management practice in terms of infiltration, water quality improvement and sustainable resource management)	**Environmental protection** (e.g., ecology, vegetation and animals); potentially a nature reserve; the original purpose might have been different	**Recreational benefits** including (water) sport, walking, fishing and bird watching; the original purpose might have been different	**Landscape aesthetics**; e.g., managed park; the original purpose might have been different

Contributor proportions (values should add up to 100):

Name of person	**Contribution (%)**

Note:
Most variables can be accurately estimated during a site visit. In case that the estimated confidence level is between 30 and 60%, it is recommended to undertake further desk study research to improve the level, and to come up with a more appropriate numerical entry. However, a confidence level entry below 30%, suggests that the variable should be treated as 'missing'.

General comments:

Appendix B: Short Survey Form for Water Bodies including SFRB

Site reference number:	**Full site name:**	**Date:**	**Template version: 041108**

Variable	% Engineered	Dam Height (m)	Dam Length (m)	Outlet Arrangement	Aquatic Animal Passage	Land Animal Passage	Floodplain Elevation (m)	Basin & Channel Connectivity	Wetness (%)	Proportion of Flow in Channel	Mean Flooding Depth	Wetness Duration	Flood Duration
Value													
Confidence													

Variable	Basin Bed Gradient	Basin Bed Flood Velocity	Wetted Perimeter	Max. Flood Water Volume	Max. Flood Water Surface Area	Mean Annual Rainfall	Drainage	Impermeable Soil %	Seasonal Influence	Site Elevation	Veg. Cover	Algal Cover	Relative Total Pollution
Value													
Confidence													

Variable	Sediment Depth	Organic Sediment Proportion	Flotsam Cover	Catchment Size	Urban	Arable	Pasture	Viniculture	Forest	Natural	GW Infiltration	Mean Depth	Length	Width
Value														
Confidence														

Variable	**Dam Condition (%)**					**Failure Hazard (%)**				**Failure Risk (%)**			
Weight (%)	30	30	20	10	10	30	35	25	10	20	50	20	10
Variable Component	Dam Structure (%)	Spillway Condition (%)	Wave Wall Condition (%)	Operation Volume Impact (%)	Other Factors Influencing Dam Condition (%)	Maximum Dam pressure (%)	Potential Loss of Life (%)	Importance of Infrastructure Affected by Dam Failure (%)	Factors Influencing Failure Hazard (%)	Dam Failure Risk (%)	Loss of Life Risk (%)	Risk of Infrastructure Failure (%)	Factors influencing failure risk
Value													
Confidence													
Rationale													
Missing Variables													
Value													
Confidence													

Bias and purpose (values should add up to 100):

Dominant hydraulic purposes	Drinking water supply	Production industry	Sustainable drainage	Environmental protection	Recreational benefits	Landscape aesthetics

Name of person	Contribution (%)		Note:
			Most variables can be accurately estimated during a site visit. In case that the estimated confidence level is between 30 and 60%, it is recommended to undertake further desk study research to improve the level, and to come up with a more appropriate numerical entry. However, a confidence level entry below 30%, suggests that the variable should be treated as 'missing'.